Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Data Deserts and Black Boxes: The Impact of Socio-Economic Status on Consumer Profiling

Nico Neumann, Catherine E. Tucker, Levi Kaplan, Alan Mislove, Piotr Sapiezynski

To cite this article:
Nico Neumann, Catherine E. Tucker, Levi Kaplan, Alan Mislove, Piotr Sapiezynski (2024) Data Deserts and Black Boxes: The Impact of Socio-Economic Status on Consumer Profiling. Management Science 70(11):8003-8029. https://doi.org/10.1287/mnsc.2023.4979

https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 70, No. 11, November 2024, pp. 8003–8029
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Data Deserts and Black Boxes: The Impact of Socio-Economic Status on Consumer Profiling

Nico Neumann,[a,*] Catherine E. Tucker,[b,c] Levi Kaplan,[d] Alan Mislove,[d] Piotr Sapiezynski[d]

[a] Melbourne Business School, University of Melbourne, Carlton, Victoria 3053, Australia; [b] MIT Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139; [c] National Bureau of Economic Research, Cambridge, Massachusetts 02138; [d] Northeastern University, Boston, Massachusetts 02115

*Corresponding author

Contact: n.neumann@mbs.edu, https://orcid.org/0000-0003-3094-6238 (NN); cetucker@mit.edu, https://orcid.org/0000-0002-1847-4832 (CET); kaplan.l@northeastern.edu, https://orcid.org/0000-0001-9136-8545 (LK); amislove@ccs.neu.edu (AM); p.sapiezynski@northeastern.edu (PS)

Received: January 19, 2021
Revised: September 8, 2022; March 23, 2023
Accepted: April 13, 2023
Published Online in Articles in Advance: January 31, 2024

https://doi.org/10.1287/mnsc.2023.4979

Abstract. Data brokers use black-box methods to profile and segment individuals for ad targeting, often with mixed success. We present evidence from 5 complementary field tests and 15 data brokers that differences in profiling accuracy and coverage for these attributes mainly depend on who is being profiled. Consumers who are better off—for example, those with higher incomes or living in affluent areas—are both more likely to be profiled and more likely to be profiled accurately. Occupational status (white-collar versus blue-collar jobs), race and ethnicity, gender, and household arrangements often affect the accuracy and likelihood of having profile information available, although this varies by country and whether we consider online or offline coverage of profile attributes. Our analyses suggest that successful consumer-background profiling can be linked to the scope of an individual's digital footprint from how much time they spend online and the number of digital devices they own. Those who come from lower-income backgrounds have a narrower digital footprint, leading to a "data desert" for such individuals. Vendor characteristics, including differences in profiling methods, explain virtually none of the variation in profiling accuracy for our data, but explain variation in the likelihood of who is profiled. Vendor differences due to unique networks and partnerships also affect profiling outcomes indirectly due to differential access to individuals with different backgrounds. We discuss the implications of our findings for policy and marketing practice.

History: Accepted by David Simchi-Levi, marketing.
Funding: Financial support from the National Science Foundation [CAREER Award 6923256] and an anonymous panel company is gratefully acknowledged.
Supplemental Material: The web appendix and data files are available at https://doi.org/10.1287/mnsc.2023.4979.

Keywords: digital advertising • marketing: segmentation • consumer profiling • algorithmic fairness • digital privacy

1. Introduction

In the age of personalized services and communications, accessing information about consumers has become an essential part of modern marketing strategy (Ansari and Mela 2003, Murthi and Sarkar 2003, Manchanda et al. 2006). To address the need for consumer information, data brokers have expanded their operations to provide detailed digital data about individuals. Data brokers like Merkle, Oracle, or Experian collect and sell valuable data about consumers to marketers, including what people buy, say, or read online.[1] Despite a rise in privacy concerns, the marketing benefits conferred by third-party data have led to flourishing demand for the services of data brokers, whose global market is forecast to grow by 6.8% annually until 2031 (TMR 2022).

Several existing studies have demonstrated a large degree of heterogeneity in the accuracy of consumer profiling, even among leading data-broker companies (Lucker et al. 2017, Neumann et al. 2019, Venkatadri et al. 2019). However, little is known about when and why audience information is more or less precise, or even available in the first place. In theory, there could be two main sources of inaccuracies in profile predictions. One source is firm-side variation, where some data brokers are better than others at accurately profiling consumers, perhaps due to proprietary technology and methods. Another explanation is consumer-side variation—that is, some consumers are just hard to profile accurately, no matter what technology is used. This paper aims to understand to what degree it is firm- or consumer-side differences that explain successful

profiling and how our findings should influence both business practice and policy.

We use five field studies to examine whether people are profiled and whether that profile is accurate for profiling for four common demographic attributes: age, state of residence, homeownership, and whether the household has children.

Our results suggest that the accuracy of profiling does not depend on different data-broker characteristics, such as reach or unique technologies.

Instead, we find that profile accuracy and coverage for demographic attributes differ significantly depending on a person's background, including race and ethnicity. Individuals with high incomes or living in more affluent areas, as well as people with a college degree, are profiled more often or have more accurate profiles. Consumers with blue-collar jobs, from single households, or who are not white are less accurately profiled. Profiling coverage depends on the same consumer characteristics, but to what degree depends on the country and the type of coverage (offline versus online).

We also present evidence that the reason that socio-economic status may affect profiling is because of differences in consumers' digital behavior. Consumers from better-off backgrounds have a broader digital footprint that data brokers can use to establish profile attributes. Those who are less well-off have a narrower digital footprint, due to owning fewer devices and spending less time online. This creates a "data desert" for those individuals, where they are less likely to have demographic profile attributes and are less accurately profiled.

Overall, our work contributes to three streams of literature.

The first is a smaller literature on the practices of data brokers and the consumer-profiling industry. Much of this research has focused on trying to quantify the value of the ability to use a tracking cookie for advertisers (Aziz and Telang 2016, Johnson et al. 2020). Other papers have documented the widespread nature of the tracking economy (Binns et al. 2018). However, only a few papers have studied the explicit actions of data brokers under the lens of consumer profiling. The exceptions are studies by Trusov et al. (2016), Neumann et al. (2019), and Venkatadri et al. (2019), who document, using different field studies, the severe biases and inaccuracies of online profile attributes. This research stream has primarily focused on how to improve probabilistic inference for provided data (Trusov et al. 2016, De Bruyn and Otter 2022). The existing studies do not consider the question of what underlying forces (demand or supply side) generally drive profile inaccuracies and coverage.

The second literature stream revolves around digital privacy. In general, the empirical literature in economics on privacy has focused on quantifying the effect on firms and technology adoption of privacy regulation (Goldfarb and Tucker 2012). This can be seen in the recent empirical privacy literature, which focuses on the consequences of General Data Protection Regulation, such as its effects on firms' measurement (Goldberg et al. 2019), market structure (Johnson and Shriver 2019), data collection (Adjerid and de Matos 2019), and venture funding (Jia et al. 2018). This paper contributes to this literature by investigating likely effects on consumers of the collection of data surrounding them in terms of its accuracy and coverage. This analysis allows insights into the differential effect of privacy regulation by people's backgrounds.

The third stream is a literature on the effect of digitization and internet technologies on inequality (DiMaggio et al. 2004). Since the early days of the internet, there have been concerns that access to the internet parallels existing sources of inequality (Keller 1995, Servon 2008). Prior work has documented the real effects of a "digital divide" in electronic commerce (Hoffman et al. 2000) and internet usage (Goldfarb and Prince 2008). Since then, there have been some efforts to try and quantify the effects of digital technologies on the "rich relative to the poor" (Miller and Tucker 2011, Tucker and Yu 2019). We contribute to this literature by being the first paper that provides empirical evidence for a relationship between the accuracy and coverage of black-box digital profiles and inequality. Although our results are based on data for marketing purposes, our finding that systemic bias in profiling affects more vulnerable consumers opens the question of the extent to which our results apply more broadly to other sectors outside of marketing (e.g., employment verification checks, financial checks, or renter background checks).

Our paper is structured as follows. Section 2 briefly describes the digital profiling and data-brokering industry. Sections 3 and 4 present Study 1 and Study 2, which use panel data sets of online users in four countries to investigate the relationship of socio-economic variables, as well as profiling accuracy and online coverage, respectively. Section 5 introduces Study 3, which examines the link of socio-economic background on profiling accuracy and offline coverage using a data set of voter-registered personal information for one of the biggest global data brokers. Sections 6 and 7 provide some evidence on the likely mechanisms underlying our results (Studies 4 and 5). Section 8 concludes with a discussion of how our results matter for marketing managers and policy makers, including possible limitations and future research directions.

2. Digital Consumer Profiling and Data Brokering

Gathering and exchanging information about consumers has been a longstanding business practice since

the mid-20th century (Levine 1995). Originally termed "database marketing," pioneers created specific lists and segments that organizations could use for direct mailing (Petrison et al. 1997). In subsequent decades, more companies specialized in building mailing lists and consumer profiles, synthesizing public records such as drivers' licenses and census data. The first professional association devoted to information brokers, the Association of Independent Information Professionals, was formed in Milwaukee in 1987 (Johnson et al. 2023). The availability of computer databases and the emergence of credit cards and the first loyalty programs allowed companies to access more personal and purchase information (Petrison et al. 1997). Thanks to the internet and digitization of products and services, consumers increasingly leave digital footprints, and the data-broker industry has experienced large growth (*The Economist* 2017).

The most common role of data brokers is to aggregate various pieces of data and then clean, analyze, and enrich the data to make them available for other organizations. Data brokers can combine information from many different sources, such as public databases and government records, web browsing records of websites, people's social media engagement, and purchase histories (FTC 2014). All the synthesized and processed consumer information is used to build profiles about people and offer these profile characteristics for profit (Johnson et al. 2023). Typically, the data are not sold directly, but, rather, licensed via subscription contracts for a particular use, such as ad targeting, risk mitigation for home loan applications, or other verification of people's credentials and background (FTC 2014, Gartner 2020). The types of information that data brokers sell encompass various forms of contact and personal information, such as full names, telephone numbers, or email addresses. Often, consumers are also segmented into similar groups of valuable characteristics, such as "homeowners," "age 35–44," or "trucking interested" (FTC 2014).

How exactly data brokers create data profiles is not public. A specific profile attribute could be drawn directly from data that the data broker compiled about the individual or inferred through algorithmic techniques. In the latter case, data brokers often use heuristics and machine learning to determine the likelihood that someone possesses a certain attribute. For example, a data broker may calculate age likelihood from a first name, ZIP code, or what car that individual drives (De Bruyn and Otter 2022). In particular, online behavior tracking through web cookies and mobile phone applications has become a major source of intelligence for data brokers. For example, if a user regularly visits retirement product websites, then that person will be classified as someone in the age range of "55+" (Trusov et al. 2016).

Sometimes, if no data about an individual can be retrieved from instances where the individual themselves revealed that data, then data brokers will try to infer the missing data from available data from other people or from other signals. For example, if one person is known to be a homeowner, data brokers often try to establish a unique pattern of behavior, such as browsing particular websites or purchasing particular items, and then assign everyone with a similar behavior to the same characteristic, such as being a homeowner (Stiebellehner et al. 2017).

3. Investigating Heterogeneity in Profiling Accuracy (Study 1)

Our research focuses on consumer profiles created by data brokers and typically sold to organizations for marketing communication. These consumer profiles are sold as prepackaged digital profile attributes, where people are classified into groups of the same characteristics for marketing purposes. These are often referred to as segments or audiences by the industry. Data brokers determine whether someone fits into the "age 25–34" or "homeowner" segments using their own methods and data sources. Organizations buy these profiles to communicate only with specific customer target groups. These practices are one of the biggest areas of data usage in the United States and grew 6.1% in 2019 (IAB and WinterberryGroup 2020).

In our study, we focus on demographic attributes for two reasons. First, these represent the most widely used attribute types bought by marketers and are offered by many different brokers (Neumann et al. 2019). Second, demographic attributes play a key role in many settings beyond marketing, such as their use in background checks for loans or risk assessments, and, therefore, are of interest for policy makers.

3.1. Method

Because the profiling methods employed by data brokers are considered proprietary information and not shared with the public, it is difficult to ascertain whether individuals' profile classifications are accurate. To establish the accuracy of a profile attribute, we used data from an international panel, where people voluntarily revealed a variety of their characteristics, and compared these with the information the data brokers claim to have established about a person's background.

We used the same method and process as Neumann et al. (2019) to validate demographic variables. For example, we compared whether the actual age provided by panelists was in the age tier that a data broker indicated for one person. The use of self-reported consumer characteristics to verify consumer profiles created by data brokers echoes existing work in marketing and information technology (Flosi et al. 2013, Lucker

et al. 2017, Neumann et al. 2019, Venkatadri et al. 2019). The validation process was conducted in 2016 and executed by syncing web cookies of panelists for four weeks with those of the digital profiles using data-management platform software.[2] There is, of course, no guarantee that our panel data are 100% accurate as ground truth, but the following considerations increase confidence in the quality of our benchmarking source. First, people's characteristics are provided on a voluntary basis, allowing individuals not to indicate any personal information if they prefer not to. Second, the panel company is International Organization for Standardization (ISO)-certified[3] for best practices and also cross-validates some information with a financial institution.

Overall, our research study allows us to benchmark the accuracy of digital profiles across four demographic attribute types (age, state of residence,[4] homeownership, family status in terms of children) for four countries: the original country in the Asia-Pacific (APAC) region (home country), another APAC country, and one country each in Europe and in America. Our country-specific samples of digital profile attributes result from two criteria:

1. The profile attributes represent widely used consumer information (Neumann et al. 2019) for which survey data were available to validate a specific digital profile attribute. Someone must have indicated whether they have children for us to be able to validate whether the family status attribute determined by data brokers is accurate.

2. Only panelists that provided this profile information were included. The desired profile information for our first two studies on socio-economic characteristics are gender and income, which were widely available in all four countries. In addition, a sample of the panelists of the home-country market have richer socio-economic information available, such as education, job type (white- versus blue-collar), and household type (family, single, or shared; multilingual versus not).

These criteria led to 3,588 individuals for the home-country sample and 5,008 for the international sample. The home-country (international) sample represents about 20% (30%) of the panelists who provided at least one variable about themselves and presents a diverse group of individuals. Web Appendix A.5 discusses in more detail questions regarding the sample representativeness. Table 1 summarizes the attributes of our validation panelists (see also Web Appendix A.1 for a correlation matrix of the seven socio-economic variables).

Across our sampled digital profiles, we were able to validate data from 15 different vendors. All possible data brokers were included for the chosen digital profile attributes, in line with our sampling criteria. We cannot reveal the identity of each data broker, but the companies in our sample represent both smaller, marketing-specific and leading, global data brokers that provide information to organizations for marketing and other purposes. We estimate that our broker sample may cover between approximately 30% and 70% of typical vendors offering consumer demographic profile data that are accessible to marketers through common data-management platforms. We discuss in Web Appendix A.9 the rationale behind this estimated range.

Table 1. Observed Socio-Economic Consumer Characteristics

Variable	Level description	Observations	Mean	SD	Median	Min	Max
Home-country sample			3,588 people				
Gender	*Male*	1,186	0.33	0.47	0	0	1
	Female	2,402	0.67	0.47	1	0	1
Income (000 s)	Continuous variable	3,588	89.79	63.40	80	5	>400
Education	*No college degree*	2,367	0.66	0.47	1	0	1
	College degree	1,221	0.34	0.47	0	0	1
Job type	*White-collar job*	3,132	0.87	0.33	1	0	1
	Blue-collar job	456	0.13	0.33	0	0	1
Homeownership	*Owns home*	2,337	0.65	0.48	1	0	1
	Rents home	1,251	0.35	0.48	0	0	1
Household type	*Family*	2,073	0.58	0.49	1	0	1
	Shared	997	0.28	0.45	0	0	1
	Single	518	0.14	0.35	0	0	1
Multilingual household	*Multilingual household*	612	0.17	0.37	0	0	1
	Main language only	2,976	0.83	0.37	1	0	1
International-country sample			5,008 people				
Gender	*Male*	2,016	0.4	0.49	0	0	1
	Female	2,992	0.6	0.49	1	0	1
Income (000 s)	Continuous variable	5,008	46.02	46.27	35	5	>400

Notes. Household status: *Single* refers to single person household; *Shared* refers to households with more than one person, but no children, such as roommates or a childless couple; *Family* refers to singles or couples who have children. Income refers to "household income before tax."

Table 2. Observed Audience Attribute Types

	Sample							
	Home country				International			
Broker	*Age*	*Family HH*	*Homeowner*	*State of residence*	*Age*	*Family HH*	*Homeowner*	*State of residence*
1	19 (4)	18 (10)	13 (7)	0	210 (56)	195 (149)	261 (197)	0
2	147 (42)	75 (48)	127 (78)	0	539 (130)	324 (233)	379 (268)	0
3	3,572 (950)	0	0	0	4,973 (985)	0	0	0
4	15 (4)	7 (3)	13 (10)	0	199 (59)	75 (62)	164 (122)	0
5	2 (0)	44 (26)	34 (21)	0	9 (2)	262 (200)	279 (214)	0
6	0	0	0	211 (154)				
7	26 (11)	26 (16)	31 (19)	0	165 (61)	208 (154)	277 (200)	63 (33)
8	282 (58)	18 (8)	18 (11)	0	856 (177)	206 (155)	242 (193)	0
9	3,724 (808)	220 (144)	201 (131)	0	5,405 (699)	508 (364)	469 (324)	50 (32)
10	4,397 (870)	807 (524)	696 (413)	3,304 (2,588)	9,272 (1,787)	2,603 (1,779)	1,082 (721)	0
11	3,585 (954)	1 (1)	1 (0)	0	5,027 (1,000)	61 (41)	87 (66)	0
12	15 (0)	19 (14)	18 (12)	0	209 (40)	273 (205)	217 (166)	0
13	487 (83)	265 (177)	37 (21)	0	1,436 (379)	1,241 (860)	1,004 (751)	0
14	4 (0)	14 (8)	7 (4)	0	111 (23)	192 (145)	180 (129)	0
15					8 (2)	0	0	0
Total per attribute	16,275	1,514	1,196	3,515	28,419	6,148	4,641	113
Correct profiles (from total)	3,784	979	727	2,742	5,400	4,347	3,351	65
Observations total	22,500				39,231			

Notes. Correct profile attributes per broker are shown in parentheses. Age and state of residence can be split up in further variants: See Web Appendix A1 for further information. HH, household.

Tables 2 and 3 summarize the audience attributes and data-broker characteristics that we study in cross-tabulation, again separately for our two samples (home country and international). Up to 14 data brokers offered digital profiles for age, whereas only 2 data brokers provided data on "state of residence." In terms of the number of observations, age is the profile attribute with the highest number (16,275 and 28,419) in both samples, whereas "Family HH" has the smallest number for the home-country sample (1,514) and "State of residence" for the international sample (113). Overall, our first study includes 4 attribute types with 15 profile attributes when considering all variants of each attribute.[5] In line with previous studies (Neumann et al. 2019, Venkatadri et al. 2019), the number of correctly classified profiles varies strongly by attribute and also by broker. For example, there were 3,784 correct profiles out of 16,275 for age and 727 correct profiles out of 1,196 for homeownership in the home-country sample. Overall, our panel samples provide 22,500 and 39,231 observations across attributes and individuals, respectively.

For our sample of data brokers, we were also able to gather data from the official data-management platform information on the number of profiles each data brokers has (*Reach*) and the types of data attributes each one offers (*Number of Segments*). The brokers in our home-country/international sample have an average of 695.6 million/595.7 million profiled online users and 13.3/13.1 audience attribute types, as shown in Tables 2 and 3, which reports our Data-Broker Covariates. Digital attributes are usually priced using fixed fees determined by the data brokers, which are then added to the remaining variable ad campaign costs by media buyers and are charged using the CPM (cost per mille) metric (Neumann et al. 2019). This means that consumer segments are sold in packages of a thousand and always at the same price. The average CPM is

Table 3. Data-Broker Characteristics

	Sample							
	Home country				International			
Data-broker covariates	Mean/median	SD	Min	Max	Mean/median	SD	Min	Max
Number of segments	13.3/15	30.0	5	17	13.1/14	2.84	7	17
Reach (users profiled [in MM])	695.6/650	471.1	200	3,500	595.7/340	376.1	200	2,000
Price in $ (CPM)	0.53/0.45	0.35	0.20	2.14	0.58/0.65	0.34	0.20	1.60

Notes. SD, Standard Deviation; CPM, Cost Per Mille.

53–58 US cents in our home-country/international sample and ranges from 20 cents to $2.14 overall.[6]

3.2. Model

In order to find out why digital profiles differ in accuracy, we first investigate the effect of consumer characteristics (the consumer side) and data-broker features (the firm side) on correct profile-attribute classifications for our sample of audience attributes. Because we have individual data on each person i and whether the classification of a data broker k is correct for a specific profile attribute j for person i, our analysis relies on binary logit regressions that link the probability of individuals being profiled correctly across data brokers and other data attributes:

$$P(CorrectProfileAttribute_{ijk} = 1) = \frac{exp^{x'_{ijk}\beta}}{1 + exp^{x'_{ijk}\beta}}, \quad (1)$$

$$x'_{ijk}\beta = \beta_0 + Attribute_j\beta_A + BrokerVariables_k\beta_B + ConsumerVariables_i\beta_C + Price_{jk}\beta_P, \quad (2)$$

where β_0 is the model's constant and $Attribute_j$ is a dummy variable for the respective audience attribute, such as "age 18–25" or "homeowner." The latter captures unobservable differences in the difficulty of profiling a certain profile attribute. For example, it may be easier to accurately assess someone being 18–24 years old versus 25–34 years old. Row vector $BrokerVariables_k$ captures the data-broker variables' effect to account for differences in their methods and abilities to create consumer profiles. These data-broker controls could be our two observable data-broker features (*Reach* and *Number of Segments*) or, alternatively, a fixed-effect dummy to control for unobservable company characteristics. $Price_{jk}$ captures the price of the audience attribute for each vendor. Row vector $ConsumerVariables_i$ captures seven socio-economic consumer characteristics for the home-country sample and can be described as follows:

$$\begin{aligned} &ConsumerVariables_i\beta_C \\ &= \beta_1 Female_i + \beta_2 \log(Income_i) + \beta_3 CollegeDegree_i \\ &\quad + \beta_4 BlueCollarJob_i + \beta_5 SharedHH_i + \beta_6 SingleHH_i \\ &\quad + \beta_7 MultiLingualHH_i. \end{aligned} \quad (3)$$

We then estimate six different model specifications (reduced and full), whereby we standardized the two numeric data-broker variables (e.g., *Reach*), while applying a log-transformation to *Income*. This reflects that log(*Income*) showed an overall better fit across our models (see also Web Appendix A.3). Given the repeated observations across individuals, we report robust standard errors clustered by individuals and attributes. As a robustness check, we also present two linear probability models, which are estimated using ordinary least squares (OLS) for the best-fitting logit model (with and without data-broker fixed effects). The standard errors of the OLS models are based on wild-bootstrapped cluster standard errors (10,000 draws), which have been suggested to improve inference for cases with a smaller number of clusters (Cameron et al. 2008, Cameron and Miller 2015).[7]

3.3. Results

Table 4 summarizes the average marginal effects (AMEs)—the change in the dependent variable for a one-unit change in an independent variable—across our eight models. Our linear regressions (OLS) naturally provide the AMEs. However, because the underlying model function is nonlinear in the case of logit models, we calculate the average marginal effects, also called average partial effects, by calculating a marginal effect for each observation unit and then averaging the result (Wooldridge 2010).

Comparing all regression results, we can see three noteworthy findings. First of all, across our estimated models that include consumer variables, we find a statistically significant relationship between our income variable and profiling accuracy. Moreover, we find statistically significant associations suggesting that individuals with blue-collar jobs or no college degree are less often profiled accurately, whereas people living in family or multilingual households have a higher chance of a correct profile for our sample.

Second, our model specifications that include segment prices (CPM) as predictors suggest either a significant negative association between price and accuracy (column (3) in Table 4) or no significant association (column (5)). Comparing data-fit criteria, such as the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and explained variances, suggests that the price variable does not provide further information that is not already captured by attribute and data-broker fixed effects.

Third, our results suggest that our data-broker variables do not seem to explain much additional variation in the data beyond consumer characteristics. The McFadden Pseudo R^2 of the logit models and the R^2 of the OLS regressions with and without the vendor variables (either fixed effects or our two continuous ones, *Reach* and *Number of Segments*) differ only by 0.001 (0.231 versus 0.232 for OLS and 0.184 versus 0.183 for logit models). The regression coefficients of the vendor fixed effects appear similar in magnitude to the consumer variables, but all except for one (vendor 7) have large standard errors. Both AIC and BIC suggest that the respective OLS and logit models without any vendor variables (columns (6) and (8)) reflect a better fit with the data.

Table 4. Average Marginal Effects: Socio-Economic Characteristics on Profiling Accuracy

Dependent variable:	*Likelihood of being profiled accurately* (mean DV = 0.366, n = 22,500)							
	(1) Logit	(2) Logit	(3) Logit	(4) Logit	(5) Logit	(6) Logit	(7) OLS	(8) OLS
Attribute FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Vendor 2		0.029	0.060	0.047			0.049	
		(0.051)	(0.042)	(0.043)			(0.049)	
Vendor 3		0.020	0.032	0.038			0.042	
		(0.053)	(0.043)	(0.045)			(0.051)	
Vendor 4		0.049	0.044	0.046			0.055	
		(0.064)	(0.054)	(0.057)			(0.070)	
Vendor 5		0.014	0.055	0.030			0.029	
		(0.048)	(0.042)	(0.042)			(0.048)	
Vendor 6		−0.025	0.154*	−0.020			−0.016	
		(0.061)	(0.066)	(0.056)			(0.060)	
Vendor 7		0.070***	0.134***	0.082***			0.092***	
		(0.020)	(0.025)	(0.019)			(0.022)	
Vendor 8		−0.001	0.037	0.013			0.009	
		(0.053)	(0.044)	(0.044)			(0.051)	
Vendor 9		0.033	0.078	0.052			0.057	
		(0.058)	(0.048)	(0.050)			(0.056)	
Vendor 10		0.032	0.018	0.050			0.056	
		(0.053)	(0.046)	(0.046)			(0.052)	
Vendor 11		0.020	0.090*	0.038			0.042	
		(0.052)	(0.042)	(0.045)			(0.051)	
Vendor 12		0.030	0.056	0.043			0.046	
		(0.088)	(0.083)	(0.086)			(0.098)	
Vendor 13		0.082	0.135**	0.097*			0.094^{+}	
		(0.053)	(0.045)	(0.046)			(0.050)	
Vendor 14		−0.034	0.013	−0.012			−0.030	
		(0.054)	(0.046)	(0.048)			(0.059)	
Number of segments					0.002			
					(0.002)			
Reach (users profiled)					−0.012^{+}			
					(0.007)			
Price (CPM)			−0.037***		−0.005			
			(0.006)		(0.003)			
log(*income*)			0.015**	0.015*	0.015*	0.015*	0.015*	0.015*
			(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Female			0.033	0.033	0.033	0.033	0.033	0.033
			(0.023)	(0.023)	(0.023)	(0.023)	(0.022)	(0.021)
College degree			0.032^{+}	0.032^{+}	0.033^{+}	0.033^{+}	0.032	0.033^{+}
			(0.019)	(0.019)	(0.020)	(0.020)	(0.020)	(0.019)
Blue-collar job			−0.022*	−0.022*	−0.022*	−0.022*	−0.021^{+}	−0.021*
			(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.010)
Shared HH			0.039	0.039	0.039	0.039	0.037	0.037
			(0.029)	(0.029)	(0.030)	(0.030)	(0.028)	(0.027)
Family HH			0.100*	0.100*	0.100*	0.099*	0.098*	0.097*
			(0.039)	(0.039)	(0.039)	(0.039)	(0.041)	(0.040)
Multilingual HH			0.046***	0.047***	0.046***	0.047***	0.048***	0.048***
			(0.007)	(0.007)	(0.010)	(0.009)	(0.012)	(0.011)
(McFadden) R^2	0.171	0.172	0.184	0.184	0.184	0.183	0.232	0.231
Log-likelihood	−12,243.9	−12,236.4	−12,057.4	−12,058.3	−12,064.0	−12,066.8		
AIC	24,517.9	24,528.8	24,186.9	24,186.7	24,178.1	24,177.7	25,115.4	25,105.9
BIC	24,638.2	24,753.4	24,475.7	24,467.4	24,378.6	24,354.1	25,404.2	25,290.3

Notes. All models have clustered standard errors (by individual and attribute). FE, fixed effects; HH, household; DV, dependent variable. OLS clustered standard errors are based on wild bootstraps (10,000 draws).
$^{+}p < 0.1$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Attribute type differences appear to account for the largest portion of explained variance. Although not shown in Table 4, the attribute fixed effects are statistically significant and show AME coefficients of high magnitude (up to 0.77; see Table A4 in Web Appendix A.7, which summarizes some examples). These results are as expected, as the fixed effects capture any unobserved differences in the difficulty to establish a

particular attribute's accuracy. We also examined a model that includes two-way interaction terms between the attribute and vendor fixed effects. The findings on the consumer covariates and price variable remain virtually the same.[8]

3.4. Extending our Results: International Sample

Next, we repeat our analysis for the international sample ($n = 39{,}321$), which comprises data from three different countries. Our intention is to demonstrate that the generalizability of our findings beyond a single sample as the level of privacy concerns and willingness to share information varies with a person's background (Zukowski and Brown 2007, Lee et al. 2019) and culture (Li 2022).

We again conduct logistic regression with the same probability as shown in Equation (1), as each individual is only represented in one country (there are no repeated observations—it's the consumer's country of residence). However, our consumer variables only include log(*Income*) and *Female* as socio-economic variables, while we include country-specific fixed-effect controls (reference category is the second APAC country) in addition to *Price* (CPM) and vendor and attribute controls. We also test a specification where we include two-way interactions between the socio-economic variables and the country fixed effects, which results in the following consumer variable equation:

$$\begin{aligned} &ConsumerVariables_i\beta_{C,int} \\ &= \beta_{1,int}Female_i + \beta_{2,int}\log(Income_i) \\ &\quad + \beta_{3,int}CountryEurope_i + \beta_{4,int}CountryAmerica_i \\ &\quad + \beta_{5,int}(Female_i \times CountryEurope_i) \\ &\quad + \beta_{6,int}(Female_i \times CountryAmerica_i) \\ &\quad + \beta_{7,int}(\log(Income_i) \times CountryEurope_i) \\ &\quad + \beta_{8,int}(\log(Income_i) \times CountryAmerica_i). \end{aligned} \tag{4}$$

We compare eight different logit model specifications and also run two OLS regressions, which are summarized in Table 5. We report again the average marginal effects and robust standard errors clustered by individuals, attributes, and countries (the OLS cluster standard errors are again based on wild bootstraps of 10,000 draws).

First, we again find that our income variable has a positive statistically significant association with profiling accuracy. Being female also has a positive association, but the precision depends on the model specification (ranging from $p < 0.1$ to $p < 0.05$). In contrast, the country fixed effects are small in coefficient magnitude and have large standard errors.

Second, in terms of prices (CPM) and accuracy, we find no significant correlation between segment price and accuracy for our tested models (columns (3) and (5)). Segment prices again provide no additional information above and beyond attribute and data-broker variables.

Third, we find that vendor-specific fixed effects increase the explained variance by only 0.004 in Pseudo R^2 (see columns (6) and (7) in Table 5).[9] This is still a small magnitude in additional explained variance, albeit slightly higher than for the home-country sample in our previous analysis. We also find that several vendor variables are statistically significant, and the best-fitting logit model according to the BIC is the model incorporating all vendor fixed effects (column (4), BIC = 39,706.8), but not the country interaction terms with the socio-economic covariates. In contrast, the AIC suggests that the best-fitting logit model is the one with vendor fixed effects and country-covariate interactions (column (7), AIC = 39,455.5). The two OLS models are in line with the logit models with respect to these two fit results (see columns (8) and (9)). However, whether we consider the models with or without interactions does not matter for the conclusions regarding the average marginal effect size.

We can only speculate as to why the vendor fixed effects appear more relevant in the international sample (versus the home-country sample). One possible reason is that the fitted models have fewer socio-economic variables that can explain accuracy differences. Instead, the latter could then be absorbed by the fixed effects for the vendors, which may have differential access to unique online users that vary in their background. We examine this in the next section.

3.5. Data-Broker Access to Different Online Users

Although vendor variables only marginally improve the explained variance in the different model specifications, there could be an indirect effect of data brokers on accuracy because each vendor has access to different people whom they can profile. We examine differential user access by comparing the socio-economic variables across our data brokers. Figure 1 summarizes differences across cookies on which brokers report for three of our key variables: income, having a blue-collar job, and having college education. The average income of users to which each broker has access varies considerably for the international sample (Figure 1(a)), but only for a few brokers for the home-country sample (Figure 1(b)).[10] Analysis of covariance confirms statistically significant differences in income across brokers for the international sample ($F(13, 24{,}324) = 13.70$, $p < 0.001$), but not the home-country one. However, we find statistically significant differences in the proportions of users with college degree (Figure 1(c), $\chi^2(13) = 93.99$, $p < 0.001$) or with blue-collar jobs (Figure 1(d), $\chi^2(13) = 27.21$, $p = 0.012$) among different data brokers for the home-country sample. These results suggest that data brokers have access to

Table 5. Average Marginal Effects: Socio-Economic Characteristics and Profiling Accuracy—International

Dependent variable:	*Likelihood of being profiled accurately* (mean DV = 0.335, n = 39,321)								
	(1) Logit	(2) Logit	(3) Logit	(4) Logit	(5) Logit	(6) Logit	(7) Logit	(8) OLS	(9) OLS
Attribute FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country interactions							Yes		Yes
Country Europe	−0.011	−0.021	0.009	0.009	0.018	0.019	0.007	0.010	0.007
	(0.016)	(0.026)	(0.028)	(0.030)	(0.027)	(0.027)	(0.040)	(0.026)	(0.032)
Country America	0.007	−0.011	−0.005	−0.005	0.012	0.012	−0.007	−0.005	−0.008
	(0.021)	(0.031)	(0.033)	(0.033)	(0.026)	(0.024)	(0.040)	(0.027)	(0.034)
Vendor 2		−0.050***	−0.051**	−0.050**			−0.050**	−0.044**	−0.044**
		(0.010)	(0.016)	(0.018)			(0.016)	(0.014)	(0.014)
Vendor 3		−0.137***	−0.134***	−0.133***			−0.133***	−0.136***	−0.136***
		(0.028)	(0.031)	(0.030)			(0.033)	(0.022)	(0.028)
Vendor 4		−0.005	−0.009	−0.008			−0.007	−0.003	−0.003
		(0.028)	(0.034)	(0.033)			(0.030)	(0.026)	(0.027)
Vendor 5		−0.005	−0.007	−0.006			−0.006	0.000	0.000
		(0.031)	(0.035)	(0.036)			(0.035)	(0.032)	(0.032)
Vendor 7		−0.021	−0.024	−0.020			−0.020	−0.014	−0.014
		(0.028)	(0.037)	(0.027)			(0.030)	(0.027)	(0.029)
Vendor 8		−0.014	−0.016	−0.014			−0.014	−0.007	−0.007
		(0.026)	(0.026)	(0.026)			(0.023)	(0.024)	(0.023)
Vendor 9		−0.099*	−0.098***	−0.096*			−0.095*	−0.091***	−0.091**
		(0.041)	(0.029)	(0.041)			(0.043)	(0.026)	(0.031)
Vendor 10		−0.091*	−0.087^{+}	−0.089*			−0.089*	−0.088**	−0.088**
		(0.037)	(0.050)	(0.042)			(0.045)	(0.029)	(0.033)
Vendor 11		−0.132***	−0.133***	−0.129***			−0.129***	−0.132***	−0.132***
		(0.031)	(0.022)	(0.032)			(0.034)	(0.024)	(0.030)
Vendor 12		−0.028	−0.030	−0.028			−0.028	−0.025	−0.025
		(0.022)	(0.018)	(0.022)			(0.019)	(0.016)	(0.017)
Vendor 13		−0.016	−0.021	−0.017			−0.016	−0.018	−0.018
		(0.033)	(0.040)	(0.036)			(0.035)	(0.025)	(0.024)
Vendor 14		−0.039*	−0.046**	−0.043^{+}			−0.043^{+}	−0.039*	−0.038*
		(0.018)	(0.015)	(0.023)			(0.023)	(0.016)	(0.017)
Vendor 15		−0.017	−0.043	−0.042			−0.045	−0.029	−0.030
		(0.160)	(0.145)	(0.146)			(0.150)	(0.117)	(0.121)
Price (CPM)			0.002		0.010				
			(0.016)		(0.018)				
Number of segments					−0.005				
					(0.007)				
Reach (users profiled)					0.000				
					(0.026)				
log(*income*)			0.041***	0.041***	0.042***	0.042***	0.041***	0.041***	0.041***
			(0.007)	(0.007)	(0.006)	(0.006)	(0.008)	(0.006)	(0.006)
Female			0.042^{+}	0.042^{+}	0.043^{+}	0.043^{+}	0.042*	0.041*	0.041**
			(0.022)	(0.023)	(0.024)	(0.025)	(0.021)	(0.016)	(0.016)
(McFadden) R^2	0.203	0.207	0.214	0.214	0.211	0.210	0.214	0.268	0.268
Log-likelihood	−19,972.1	−19,866.5	−19,699.9	−19,700.0	−19,789.1	−19,800.0	−19,694.8		
AIC	39,972.3	39,787.0	39,459.8	39,458.0	39,616.2	39,631.9	39,455.5	40,325.6	40,324.9
BIC	40,092.4	40,018.6	39,717.2	39,706.8	39,779.2	39,769.2	39,738.7	40,583.0	40,616.6

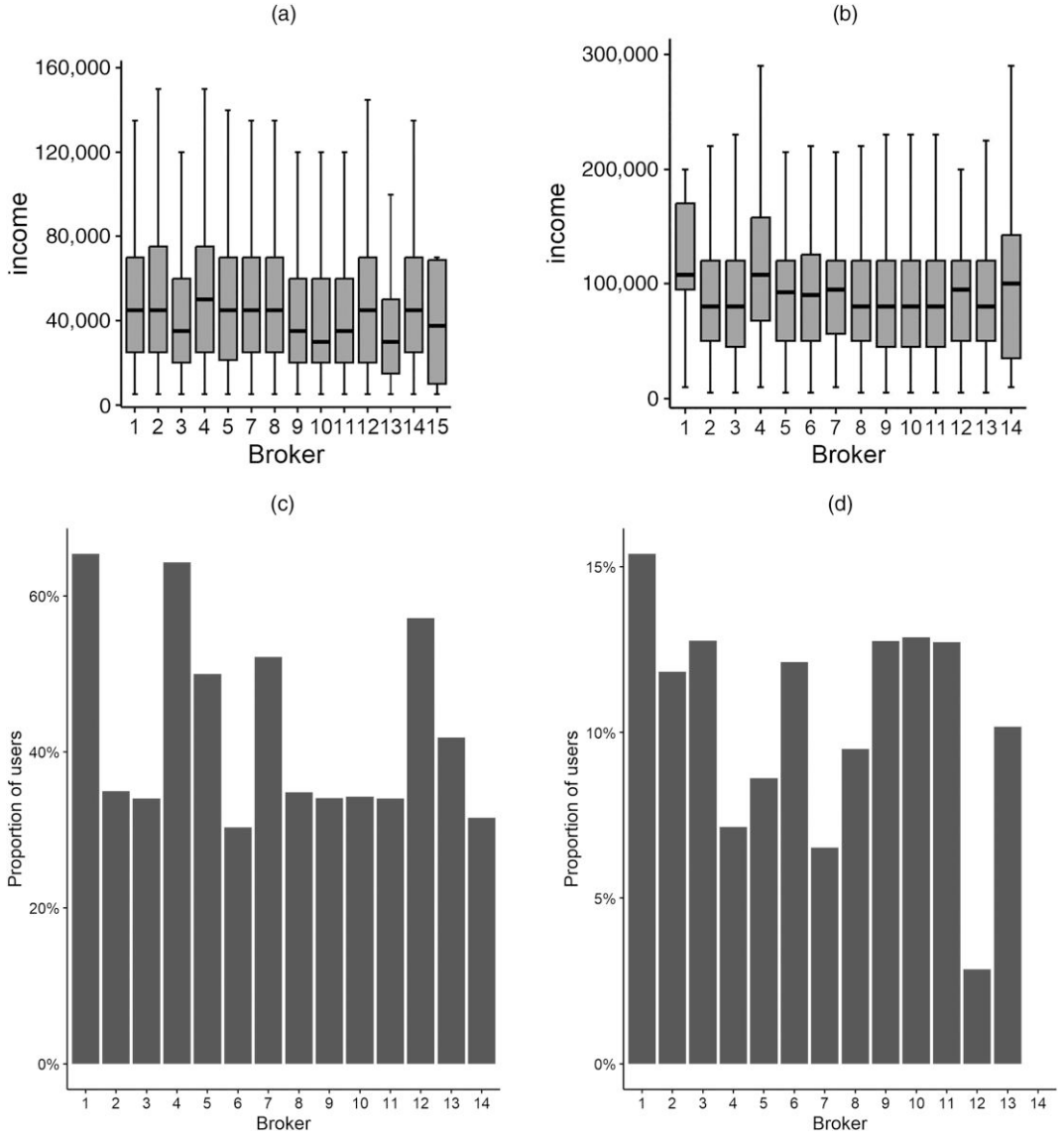
Notes. All models have clustered standard errors (by individual and attribute). FE, fixed effects; DV, dependent variable. OLS clustered standard errors are based on wild bootstraps (10,000 draws).
$^{+}p < 0.1$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

a different pool of people in terms of socio-economic status.

4. What Drives Online Coverage (Study 2)

Our previous analysis found that the accuracy of profile attributes depended on who was being profiled. We now turn to investigate whether socio-economic characteristics may influence who is being profiled online. Individual data brokers do not reveal their decisions on who is classified and who is not in data-management platforms for online campaigns. Therefore, we do not have data on people who are not classified when relying on audience segments provided by

Figure 1. Differences in Key Consumer Characteristics Across Data Brokers



Notes. (a) Income average (international). (b) Income average (home country). (c) Users with college degree. (d) Users with blue-collar jobs.

data-management platforms. We only observe how data brokers have classified a cookie or user for a particular attribute. However, this is a practical limitation affecting every brand that would like to use consumer profiling for online marketing campaigns. Study 2 therefore presents a field test examining what we call *online* coverage. We investigate whether specific users have been profiled more often across the accessible websites of our 15 different data brokers for our test period, which is also a common length for advertising campaigns. Methodologically, we perform the following analysis: although we have repeated observations of users across data brokers, not every data broker profiled every user. We remind the readers that data vendors differ in their networks, and not every broker will have the ability to profile the same users. We can use this variation in profiling frequency across users in our two samples (home country and international) to obtain possible evidence on how socio-economic covariates affect online profiling coverage in a field test mimicking an ad campaign.

4.1. Method and Model

To empirically investigate the probability of "being profiled online," we create a new data set with every user and data-broker combination for our samples and then mark which user was profiled by a certain vendor. We obtain 50,232 observations (14 data brokers × 3,588 individuals) for the home-country sample and 70,112 observations (14 data brokers × 5,008 individuals) for

the international sample. We then again perform logistic regressions and separate analyses for the home-country and international samples, which differ in the available socio-economic variables. The model specification are similar to the analysis in Study 1, but the dependent variable is the likelihood to be *profiled* (given the covariates), and the focus is on the user-data broker combinations only:

$$P(UserProfiled_{ik} = 1) = \frac{exp^{x'_{ik}\beta_{Cov}}}{1 + exp^{x'_{ik}\beta_{Cov}}}, \quad (5)$$

$$x'_{ik}\beta_{Cov} = \beta_{0,cov} + BrokerVariables_k\beta_{B,Cov} + ConsumerVariablesCoverage_i\beta_{D,Cov}, \quad (6)$$

where we include the same $BrokerVariables_k$ and $ConsumerVariablesCoverage_i$ as in our analysis on profiling accuracy. That is, we include either data-broker fixed-effect controls or the two continuous data-broker variables and the seven/two socio-economic characteristics for the home-country/international sample. We will also test again interaction terms between log(*Income*)/*Female* and country fixed effects. For our coverage analysis, we further include the age of a user as an additional covariate. We were not able to include *Age* in Study 1, as it is also a dependent variable in the accuracy analysis. By contrast, in Study 2, the dependent variable is whether a user was profiled for one broker or not. The results of seven different logit models and two OLS regressions are summarized in Table 6. We carry out this analysis at the data-broker level, which means that a user counts as profiled if they are classified for any of our attributes by one vendor. The data are too sparse, unfortunately, to analyze attribute-specific coverage.

4.2. Results

Our online coverage analysis finds that, first, in contrast to our profiling-accuracy analysis, data-broker characteristics matter. The AME coefficients for the broker fixed effects are overall much larger (up to 0.992 in absolute values), and many are estimated with higher precision, reaching statistical significance. We also find statistically significant positive associations between the number of segments that a broker offers and coverage for both samples (columns (2) and (6) in Table 6) and for broker reach and coverage in the home-country sample (column (2)). However, again, the models with data-broker fixed-effect controls, which also account for unobservable characteristics, still fit the data better according to AIC and BIC (comparing columns (2) and (3) or columns (6) and (7)). Although AIC and BIC are in agreement about which model is the overall best-fitting one for the home country (column (3)), this is not the case for the international sample. For the latter, the AIC suggests the model with interaction terms is the best-fitting one (29,868.6 versus 29,881.1), whereas the BIC points to the model without interactions (30,055.1 versus 30,097.6). Socio-economic variables or vendor effects seem to matter less. Most importantly, adding vendor fixed effects to the logit [OLS] model increases the explained variance ((McFadden) R^2) strongly—that is, from virtually 0% to 79.0% [84.6% for OLS] for the home country (see columns (1) and (3)) and from 2.8% to 67.1% [69.8% for OLS] for the international sample (see columns (5) and (7)/(8) for the logit models or columns (4) and (9) for the OLS regressions).

Second, we find fewer associations between socio-economic variables and the likelihood of being profiled across data brokers, with strong differences across our two different samples (and countries). For the home-country sample, there are only two variables that correlate with the likelihood of being profiled (columns (1)–(4)). Users with a college degree are more likely to be profiled. People living in a multilingual home are less likely to be profiled. For the international sample, we find consistent significant positive correlations of income with being profiled, whereas some models also suggest a positive relationship between profiling coverage and being female. The logit model including interactions of our socio-economic covariates (age, female, income) and country fixed effects also suggests an overall positive correlation of age and profiling coverage. Although presenting the coefficient results as AMEs helps for interpretation purposes, they do not show how the interactions influence signs and coefficients of the country interaction terms (which are averaged out for the AME calculations). In Web Appendix A.8, we present the logit model results as log odds, shedding light on the coefficient signs and magnitudes of the actual interaction terms. The individual interaction results show that income has a positive correlation to coverage in all countries, whereas the sign and magnitude of age and gender coefficients vary across countries for our sample. This finding seems reasonable, as socio-economic status will vary with age and gender across nations. The two OLS regressions, based on the best-fitting logit models for each sample according to the BIC, demonstrate that our general findings do not depend on the functional form or type of clustering standard errors.

4.3. Online Marketing Campaigns, Selection Effects, and Identity Fragmentation

Overall, Study 2 provides evidence for the role of socio-economic characteristics for the likelihood of being profiled online by data brokers in a marketing platform. Our results also suggest some differences as to which consumer variables matter across countries. This is not surprising because cultural differences and local market

Table 6. Average Marginal Effects: Socio-Economic Characteristics and Online Coverage

Dependent variable:	*Likelihood of being profiled across data brokers*								
	Home country (mean DV = 0.306, n = 50,232)				*International* (mean DV = 0.347, n = 70,112)				
	(1) Logit	(2) Logit	(3) Logit	(4) OLS	(5) Logit	(6) Logit	(7) Logit	(8) Logit	(9) OLS
Country Europe					0.053***	0.053***	0.067***	0.067***	0.051***
					(0.002)	(0.002)	(0.001)	(0.001)	(0.002)
Country America					0.270***	0.271***	0.232***	0.230***	0.291***
					(0.001)	(0.001)	(0.013)	(0.012)	(0.000)
Country interactions								Yes	
Vendor 2			0.066***	0.045***			0.039**	0.039**	0.043*
			(0.006)	(0.004)			(0.012)	(0.012)	(0.017)
Vendor 3			0.338***	0.988***			0.534***	0.534***	0.913***
			(0.010)	(0.002)			(0.043)	(0.043)	(0.079)
Vendor 4			0.002	0.001			−0.021*	−0.021*	−0.018
			(0.005)	(0.001)			(0.009)	(0.009)	(0.019)
Vendor 5			0.026***	0.009***			−0.004	−0.004	−0.004
			(0.006)	(0.002)			(0.019)	(0.019)	(0.016)
Vendor6			0.068***	0.048***					
			(0.006)	(0.004)					
Vendor 7			0.019***	0.006***			0.005	0.005	0.005*
			(0.005)	(0.002)			(0.005)	(0.005)	(0.002)
Vendor 8			0.072***	0.054***			0.051***	0.051***	0.060*
			(0.006)	(0.004)			(0.012)	(0.012)	(0.026)
Vendor 9			0.345***	0.989***			0.546***	0.546***	0.914***
			(0.011)	(0.002)			(0.050)	(0.050)	(0.078)
Vendor 10			0.231***	0.887***			0.281***	0.281***	0.669***
			(0.006)	(0.005)			(0.018)	(0.018)	(0.088)
Vendor 11			0.393***	0.992***			0.647***	0.646***	0.919***
			(0.020)	(0.001)			(0.047)	(0.047)	(0.079)
Vendor 12			0.010^{+}	0.003^{+}			0.001	0.001	0.001
			(0.005)	(0.001)			(0.005)	(0.005)	(0.003)
Vendor 13			0.108***	0.157***			0.178*	0.178*	0.343*
			(0.006)	(0.006)			(0.087)	(0.087)	(0.172)
Vendor 14			−0.010	−0.002			−0.032*	−0.032*	−0.026
			(0.007)	(0.001)			(0.013)	(0.013)	(0.027)
Vendor 15							−0.288*	−0.288*	−0.079
							(0.128)	(0.128)	(0.078)
Reach (users profiled)		0.068***				0.051			
		(0.001)				(0.047)			
Number of segments		0.156***				0.145***			
		(0.001)				(0.020)			

Table 6. (Continued)

Dependent variable:	*Likelihood of being profiled across data brokers*								
	Home country (mean DV = 0.306, n = 50,232)				*International* (mean DV = 0.347, n = 70,112)				
	(1) Logit	(2) Logit	(3) Logit	(4) OLS	(5) Logit	(6) Logit	(7) Logit	(8) Logit	(9) OLS
log(*Income*)	0.001	0.001	0.001	0.001	0.009*	0.009*	0.008***	0.008***	0.009*
	(0.001)	(0.001)	(0.001)	(0.001)	(0.004)	(0.004)	(0.002)	(0.000)	(0.004)
Age	0.001	0.001	0.001	0.001	0.002	0.002	0.001	0.001***	0.002
	(0.001)	(0.001)	(0.001)	(0.001)	(0.003)	(0.003)	(0.003)	(0.000)	(0.003)
Female	−0.001	−0.001	−0.001	−0.001	0.011^{+}	0.011^{+}	0.011**	0.011***	0.011*
	(0.003)	(0.003)	(0.003)	(0.003)	(0.006)	(0.006)	(0.004)	(0.000)	(0.005)
College degree	0.009***	0.009***	0.009***	0.009***					
	(0.003)	(0.003)	(0.003)	(0.003)					
Blue-collar job	−0.003	−0.003	−0.004	−0.003					
	(0.003)	(0.003)	(0.003)	(0.003)					
Shared HH	−0.002	−0.002	−0.002	−0.002					
	(0.004)	(0.004)	(0.004)	(0.004)					
Family HH	−0.002	−0.002	−0.002	−0.002					
	(0.003)	(0.003)	(0.003)	(0.003)					
Multilingual HH	−0.009**	−0.009**	−0.010**	−0.009**					
	(0.003)	(0.003)	(0.003)	(0.003)					
(McFadden) R^2	<0.001	0.079	0.790	0.846	0.028	0.088	0.670	0.671	0.698
Log-likelihood	−30,914.3	−28,488.5	−6,504.2		−43,980.6	−41,267.6	−14,921.6	−14,909.3	
AIC	61,846.5	56,999.0	13,052.4	−29,235.2	87,973.3	82,551.1	29,881.1	29,868.6	11,028.6
BIC	61,926.0	57,096.0	13,246.6	−29,032.2	88,028.2	82,624.4	30,055.1	30,097.6	11,211.8

Notes. All models have clustered standard errors (by individual). FE, fixed effects; HH, household; DV, dependent variable. OLS clustered standard errors are based on wild bootstraps (10,000 draws).

$^{+}p < 0.1$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

influences can lead to varying consumer behavior or relevance of socio-economic status.

However, although mimicking an online ad field test, we need to consider that Study 2 investigates strictly online coverage that can be subject to different selection effects and biases, some of which are unavoidable for any digital campaign relying on web cookies. Specifically, our results depend on the following three factors:

1. The partner network of the panel provider and the online activity of panelists who revealed the relevant information we needed for this research.

2. The online activity of users and the partner network that a data broker can access for their profiling method on selected partner websites (Lewis and Reiley 2014).

3. Data brokers and the platform providers ensuring the correct matching of online users on different websites and networks via web cookies, which can be challenging, given that individuals tend to have multiple browsers and devices (Lin and Misra 2022).

Our use of multiple samples from different countries addresses the first possible selection effect that stems from our use of surveys to evaluate the ground truth. In contrast, the presence of online activity bias or identity-matching errors are unavoidable selection effects in any online campaign or measurement effort (Lin and Misra 2022). Therefore, Study 2 can be interpreted as reflecting the online coverage outcome that a brand can achieve in an digital ad campaign relying on demographic attributes. However, it cannot distinguish the relative role of a broker's lack of profile offering for some people relative to other kinds of bias.

In our next study, we address this by moving to an offline context, where these last two selection effects are not present.

5. Profiling Accuracy and Offline Coverage (Study 3)

Studies 1 and 2 investigate profiling accuracy and coverage using an online panel that is integrated with a data-management platform, which serves as a connection hub to a large number of data brokers. This unique setup is the most common way to buy profile attributes for online campaigns for marketers, but only allows examining the final outcome of coverage, rather than distinguishing between different types of selection effects and biases.

In our third study, we investigate profiling coverage and accuracy based on a complementary, but different, methodology that aims to control for any selection effects and possible process errors as much as possible. First, we use publicly available voter records from North Carolina as the validation source of consumer characteristics. We deem these data as reliable ground truth, as the respective attributes are self-reported by individuals who face potential criminal charges for misrepresentation.

Second, instead of matching users online via web cookies (which is one likely error source in any online campaign), we upload our test data directly to a data broker's online interface using individuals' addresses and full names as matching criteria. We remove a known attribute from our consumer file and let the data broker then append the missing demographic information—in this case, "age." We then attempt to purchase the age information about each individual from the data broker. In other words, we let the data broker append the missing information on age. We outline the details of our methodology, the model, and the results in the subsequent sections.

5.1. Method

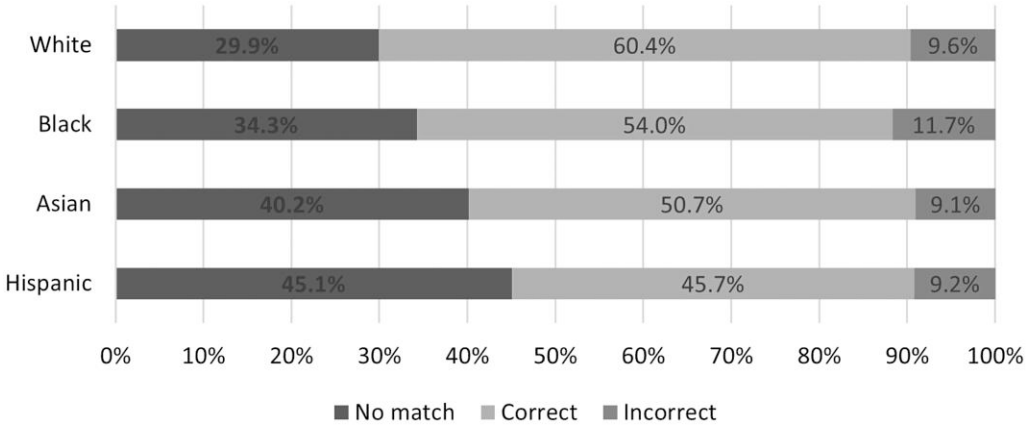
In total, we use three separate data sources for our third study: North Carolina voter records from 2021, a web interface (data append), and the U.S. Census. We use the self-reported voter records as a ground truth for the following consumer characteristics: ethnicity/race, age, gender, name, and address of individuals. These records are published on the web by the North Carolina State Board of Elections and released every week, thus keeping reported characteristics in the voter records up-to-date.[11] For ethnicity/race, we distinguish four categories in line with the U.S. Census categories and voter records: Hispanic, Asian, African American (referred to as Black), and white.[12]

For our study, we create a stratified sample and select 8,930 voters with up-to-date registration from each of our four race and ethnicity groups (Hispanic, Asian, Black, and white individuals). Our choice of individuals is random, with the constraint that each age range has the same number of individuals in each racial group (for age ranges 18–24, 25–34, … , 85–94); within each race/age combination, the number of men and women of each race are equal.

Unfortunately, the voter data do not provide information regarding the socio-economic status of each person. Instead, we create a ZIP-code-level covariate "poverty level" based on the Census data as a proxy.[13] Specifically, we used the 2019 American Community Survey (ACS) to obtain information about the poverty rates in each ZIP code where any individual in our data set resides. After discarding 456 people who indicated post office (PO) boxes as mailing addresses, we obtain a final sample of 35,264 (Asian: 8,798; Black: 8,802; Hispanic: 8,831; white: 8,833).

For our validation test, we remove the actual provided age from our file and then upload the data to the web interface of a globally leading data broker. We match the consumers using their name and address and purchase for $0.015 per record the age information

Figure 2. Age Coverage and Accuracy for Voters of Different Ethnicity/Race



for each one. Although we cannot reveal the identity of the broker we used, the vendor is one of the largest in the world and provides data services to multiple industries (and not only marketers).

The data broker then returns the following information for each person in our list: the type of match (Non-Match, Geographic Match, Household Match, and Person Match) and the certainty of the purchased attribute—in our case, birth age (None, Estimated, Exact). For our analysis presented here, we treat anything other than a person match as not matching and only verify the correctness of the exact matches. For the latter, we define any deviation of more than one year (plus or minus) from the self-reported birth age as incorrect. Although not reported here, we highlight that our results are robust to various relaxations of these definitions.

We calculate the rates of matches and correct age classifications per race and summarize the results in Figure 2. This descriptive analysis reveals some differences among races and ethnicities, both in terms of coverage and accuracy. Hispanic voters have the highest rate of failing to match based on name and street address at 45.1%, compared with white voters at 29.9%. Furthermore, there are some notable disparities in accuracy, even for the individuals for whom a match is available. For example, the age of 11.7% of Black voters is wrong by more than a year, compared with 9.1% of Asian voters.

Although suggestive, our summary analysis by ethnicity/race does not account for the influence of gender or ZIP-code poverty. We therefore perform binary choice model analysis again to better understand the average marginal effects of each covariate on profiling coverage and accuracy.

5.2. Model

Our model specification is similar to the setup in Studies 1 and 2, but we have one observation per individual and only consumer characteristics as covariates. We use a bivariate probit model specification to model the joint likelihood of being profiled and having an accurate profile attribute (in this case, "age") given the covariates as follows:

$$P(AgeMatch_i = 1) = P(\alpha_S + ConsumerVariablesSelect_i\beta_S + u_i > 0), \quad (7)$$

$$P(CorrectAge_i = 1) = P(\alpha_O + ConsumerVariablesOutcome_i\beta_O + v_i > 0), \quad (8)$$

where $CorrectAge_i$ is only observed if $AgeMatch_i = 1$. Row vectors $ConsumerVariablesSelect_i$ and $ConsumerVariablesOutcome_i$ capture the consumer characteristics for the selection and outcome equations, respectively, and error terms (u_i, v_i) are jointly distributed as standard bivariate normal with zero means and correlation coefficient ρ. This Heckman-style bivariate probit model controls for endogenous sample selection bias, which persists in case ρ is different from zero (Heckman 1979, Van de Ven and Van Praag 1981).

We carry out three different specifications of our model. The consumer variables in the Selection Equation (7) ($ConsumerVariablesSelect_i$) include an indicator variable for women (*Female*), race (*Asian*, *Black*, *Hispanic*), ZIP-code poverty (*Poverty*, log-transformed),[14] birth age (*Age*), whether someone has a driver's license (*Driver*), and the political party membership status (*Party 2*, … ,*5*). Because we are interested in the effect of belonging to a certain ethnic minority and ZIP-code poverty, we carry out reduced and full model specifications for the Outcome Equation (8). Specifically, we include either all consumer variables (Model A), omit *Poverty* (Model B), or omit the three ethnic indicators (Model C). We omit either of these variables, as these capture similar background characteristics. It should be noted that the outcome equation does not include *Age*

Table 7. Average Marginal Effects: Bivariate Probit on Profiling Coverage and Accuracy

	(1) Model A	(2) Model B	(3) Model C	(4) Model A	(5) Model B	(6) Model C
Variable	Accuracy (outcome equation)			Coverage (selection equation)		
log(*Poverty*)	−0.00789+		−0.0116**	−0.0425***	−0.0435***	−0.0420***
	(0.00425)		(0.00397)	(0.00484)	(0.00481)	(0.00484)
Female	0.0345***	0.0338***	0.0333***	−0.0167***	−0.0167***	−0.0167***
	(0.00432)	(0.00422)	(0.00421)	(0.00504)	(0.00504)	(0.00504)
Age				0.00610***	0.00610***	0.00610***
				(0.000152)	(0.000152)	(0.000152)
Black	−0.0391***	−0.0394***		−0.0423***	−0.0421***	−0.0474***
	(0.00704)	(0.00691)		(0.00811)	(0.00811)	(0.00822)
Hispanic	−0.0185*	−0.0169*		−0.149***	−0.149***	−0.151***
	(0.00745)	(0.00721)		(0.00723)	(0.00723)	(0.00718)
Asian	−0.0109	−0.00798		−0.115***	−0.115***	−0.116***
	(0.00681)	(0.00644)		(0.00731)	(0.00731)	(0.00727)
Driver	−0.00632	−0.00626	−0.00609	0.0583***	0.0582***	0.0583***
	(0.00729)	(0.00713)	(0.00716)	(0.00819)	(0.00819)	(0.00819)
Party 2	0.00312	0.000126	0.00763	0.249**	0.249**	0.248**
	(0.0981)	(0.0958)	(0.0958)	(0.0872)	(0.0872)	(0.0872)
Party 3	0.0147	0.0113	0.00167	0.272**	0.272**	0.273**
	(0.0981)	(0.0958)	(0.0957)	(0.0871)	(0.0871)	(0.0871)
Party 4	−0.127	−0.125	−0.126	0.0890	0.0892	0.0894
	(0.138)	(0.135)	(0.135)	(0.142)	(0.142)	(0.142)
Party 5	−0.0189	−0.0202	−0.0191	0.220*	0.220*	0.220*
	(0.101)	(0.0988)	(0.0989)	(0.0921)	(0.0921)	(0.0921)
Party 6	0.00575	0.00330	0.00357	0.267**	0.267**	0.267**
	(0.0980)	(0.0957)	(0.0957)	(0.0871)	(0.0871)	(0.0870)
ρ	−0.284***	−0.319***	−0.314***	−0.284***	−0.319***	−0.314***
	(0.0727)	(0.0710)	(0.0626)	(0.0727)	(0.0710)	(0.0626)
n		22,090			35,264	
Mean DV		0.842			0.626	
Log likelihood	−31,755.8	−31,757.6	−31,774.2	−31,755.8	−31,757.6	−31,774.2
AIC	63,563.6	63,565.2	63,594.4	63,563.6	63,565.2	63,594.4
BIC	63,783.8	63,777.0	63,789.2	63,783.8	63,777.0	63,789.2

Notes. All models have robust standard errors. DV, dependent variable.
$^{+}p < 0.1$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

for identification purposes. We summarize the results of our three Heckman-style probit regressions in Table 7.

5.3. Results

First, we note that our correlation coefficient ρ is around −0.3 across our three model specifications and is statistically significant. This finding implies that our bivariate probit structure is required to obtain unbiased estimates for our parameters.

Second, our bivariate probit model shows that all our consumer characteristics have a significant effect on offline profiling coverage. Women, people living in poorer areas, and people without a driver's license are less likely to have a match with exact age. Hence, the role of poverty (which can be seen as a proxy for income, albeit at the ZIP-code level) appears to matter more for offline coverage than for online coverage (as shown in Study 2), where the association was more varying across our two samples. Moreover, Black, Hispanic, and Asian individuals are less likely to have exact age information, with being Hispanic having the greatest marginal impact (being about 14.9% less likely to be profiled versus a white individual) from the three race variables. We also find that a person's age has a positive correlation with profiling coverage in our bivariate probit. This was the case for some model specifications in Study 2—similar to the gender variable—whereby the sign, magnitude, and precision seem to depend on the country and context (recall that age/gender has significant interactions with the country in Study 2). The party membership covariates, as well as having a driver's license, have significant positive associations with profiling coverage as well. We believe this is very plausible, as any kind of information that is available in potential databases (such as electronic entries about political membership or driver's license) allows data brokers to create a profile.

Conditional on being successfully matched, the average error rate is 15.8%, which appears lower than the rate we observed for our online marketing context in Study 1. Despite the overall lower error rate, the bivariate

probit still suggests several statistically significant relationships between covariates and profiling accuracy. For example, women are significantly more likely to be accurately profiled (about 3.4%) than men. This was the case in Study 2 as well, although the precision of the gender variable was much greater for the international sample (implying that it may vary by context and country, too). We further find in Study 3 that Hispanic and Black voters are less likely to be profiled accurately (about 1.7%–1.9% and 3.9% less in comparison with white voters, respectively; see columns (1) and (2)).

With respect to the poverty of an area where a person lives, we find a negative correlation with profiling accuracy, although the coefficient of log(*poverty*) only reaches greater precision (at the traditional statistical significance level of $p < 0.05$) if we don't include people's ethnicity/race in the model (see columns (1) and (3)). One likely reason is that a person's ethnicity/race may partially capture similar effects to the poverty of their area. For example, Black voters tend to live in area codes with higher poverty (see Web Appendix A.4). However, even for the model without ethnicity/race (in column (3) of Table 7), the magnitude of the log(*poverty*) coefficient is smaller for the accuracy than for the coverage outcome. Hence, although the identified variable-outcome relationship is in agreement with the results of Study 1, the poverty variable appears not as prevalent for accuracy as for coverage in Study 3.

Finally, we find no significant associations of having a driver's license or political party membership with profiling accuracy. Hence, these consumer variables may help data brokers create a profile (as suggested by the coverage analysis), but not establish accurate profile information. The latter is likely to depend on the scope and intensity of available signals and the profile classification methods used by data brokers, as we will explore in the next section.

6. Digital Footprint and Profiling Accuracy and Online Coverage (Study 4)

In Studies 1–3, we present evidence for the role of socio-economic status for consumer profiling accuracy and coverage of demographic attributes. One remaining question is what the underlying mechanism(s) for the identified relationships could be. Because data brokers require access to some kind of accessible information to be able to profile a consumer, it stands to reason that the digital footprint of individuals is likely to affect both profiling coverage and accuracy. We examine the relevance of different digital behavior for our context in our fourth study.

6.1. Method and Model

We return to the panel provider we used for Study 1 and Study 2 and obtain relevant measures on three measures for digital behavior for a sample of 4,111 panelists across the same four countries as in Study 1 and Study 2. We are able to retrieve the following information that was provided again directly by panelists:

1. Whether someone shops online (yes, no).
2. The number of electronic devices a person owns (e.g., TV, computer, tablet, gaming console, etc.).
3. The number of regular online and social media activities in which a person engages (e.g., sharing news content, using messengers, sharing photos on social media, using online dating, uploading videos, blogging, video gaming, etc.).

Table 8 provides a descriptive summary for the three digital-footprint measures for our sample. We find that 90% of users shop online. Individuals, on average, engage regularly in 3.6 online activities and own on average about seven electronic devices.

We then investigate the effect of users' digital-behavior characteristics on profiling accuracy and online coverage using the same approaches as in Study 1 and 2, but with a different sample[15] and our three consumer characteristics only (i.e., *Devices*, *Activities*, and *Online Shopping*). Because we investigate the same covariates now for our home-country sample and the three other countries, we analyze accuracy and coverage using data from all four countries. Specifically, we again examine the likelihood of being accurately profiled across our four demographic attributes and online coverage across our sample of 15 data brokers for the four-country sample. For both coverage and accuracy, we perform various logistic regressions using full and reduced model specifications, including a specification with (two-way) interactions of country and our three digital covariates. For example, we include either vendor fixed effects, our two continuous vendor variables (*Reach* and *Number of Segments*), or no vendor controls. Because our three digital-behavior variables of interest capture similar consumer traits (e.g., people who own many devices are

Table 8. Observed Consumer Characteristics for Online Behavior

Covariate	Level description	Mean	SD	Median	Min	Max
Online Shopping	Buys goods and services online	0.904	0.294	1	0	1
Online Activities	Count variable	3.609	2.755	3	1	15
Devices	Count variable	7.205	2.216	7	0	13

also more likely to engage in other online activities, $r(1,265) = 0.33$ with $p < 0.001$), we investigate their impact jointly and individually. As a robustness check, we perform an OLS regression of the best-fitting logit models. All models have cluster-adjusted robust standard errors again, including wild bootstrapping for OLS.

6.2. Results

Tables 9 and 10 summarize the average marginal effects for digital-behavior characteristics on profiling accuracy and coverage, respectively. For the former, we find a significant positive association of all three digital-behavior variables across our different model specifications (see columns (8)–(10) in Table 9). However, coefficient magnitude and precision of the three digital variables differ. Two variables stand out across all regressions. First, *Online Shopping* has the greatest average marginal effect overall (the accuracy increases by 3.4%–3.5% for people who shop online, without considering other variables, or 1%–1.2% conditional on the number of devices and online activities of a person). Second, *Devices* appears to be the most precisely estimated covariate and the only one reaching statistical significance if we include all three digital variables in a model (see columns (1)–(4)). Our estimated models suggest that profiling accuracy increases by 0.8%–1.2% for each additional device someone owns.

Another important observation is that the three covariates change both in precision and magnitude when including the covariates jointly and individually. For the latter, they all increase their magnitude and reach statistical significance (albeit with highest precision for our OLS models). This finding suggests that all three of them capture to some degree similar traits, as we suspected. However, despite the three variables capturing partly similar digital consumer behavior, the best-fitting model is still one with all three covariates included. Which one is the best-fitting model specification overall depends on the criterion. The BIC points to the model with only our two continuous data-broker variables (column (2), BIC = 30,889.3) and the AIC to the one with data-broker fixed effects (column (3), AIC = 30,634.6).

Regarding profiling [online] coverage, we find a similar pattern across our three digital covariates, but also some key differences in terms of overall impact (i.e., the magnitude of associations). *Online Shopping* still has the largest coefficient (between 2.3% and 5.2% increases in average marginal effects), but shows a large variance across nearly all models, too. Similar to the accuracy analysis, *Devices* stands out in terms of its precision, as it is the only covariate reaching statistical significance across multiple models: the probability to be profiled by a broker in our sample increases by about 0.16%–0.18% for each additional device someone owns.

With regard to overall influences of our different variables, we note the following three findings. First, the magnitude of the coefficients of our three digital consumer covariates is much smaller in our coverage regressions than in the accuracy regressions, even by factors of two or more. For example, considering the OLS regressions in columns (8)/(9)/(10), the coefficient point estimates are 0.034/0.009/0.012 for accuracy versus 0.0045/0.002/0.0017 for coverage, respectively. This finding reveals that our three observed digital consumer behavior variables matter more for accuracy than coverage in our data.

Second, we find larger standard errors in the covariate coefficients for the online coverage regressions in Table 10 (in comparison with the accuracy regressions in Table 9). The greater variation in covariate estimates for coverage can also be seen by the fact that the model including country-covariate interaction terms (column (4), Table 10) is the only one suggesting significant average marginal effects for all covariates (although it does not have the overall best data fit). Hence, for profiling coverage, there seem to be stronger differences in terms of country-specific effects than for the accuracy analysis. Here, we remind the reader to interpret the results with caution, given that the data, in particular for online coverage, are likely subject to biases that are typical of digital marketing campaigns, such as identity fragmentation (Lin and Misra 2022).

Third, even though our digital covariates appear to play a smaller role for profiling coverage than for profiling accuracy, the data-broker variables appear to be much more important variables for the former. Although adding observable vendor characteristics to the model explains little variation for our accuracy models (see columns (1) and (3), Table 9), the McFadden Pseudo R^2 increases from 1.9% to 70.5% or higher in the case of our coverage models (see columns (1) and (3), (4), (5), (6), and (7) in Table 10). For OLS regressions, the explained variance is even 74.8% for profiling coverage (see columns (8), (9), and (10) in Table 10). The result of differential relevance of data-broker characteristics for profiling accuracy versus coverage is also in line with our analyses in Studies 1 and 2 (where we achieved explained variances of up to 79%/84.6% for coverage regressions using logit/OLS models, respectively). As discussed earlier, we argue that profiling coverage depends much more strongly on unique networks and partnerships of data brokers.

7. Linking Digital Footprint and Socio-Economic Characteristics (Study 5)

We have shown in Studies 1–4 that socio-economic characteristics, as well as consumers' digital footprint, affect both profiling accuracy and coverage for demographic attributes. We posit that these two consumer-

Table 9. Average Marginal Effects: Digital-Behavior Characteristics and Profiling Accuracy

Dependent variable:	*Likelihood of being profiled accurately* (mean DV = 0.341, n = 30,513)									
	(1) Logit	(2) Logit	(3) Logit	(4) Logit	(5) Logit	(6) Logit	(7) Logit	(8) OLS	(9) OLS	(10) OLS
Attribute FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country interactions				Yes						
Country Asia-Pacific	0.020^{+}	0.020^{+}	0.021^{+}	0.021	0.021*	0.022*	0.021*	0.020*	0.021*	0.020*
	(0.011)	(0.011)	(0.011)	(0.020)	(0.011)	(0.010)	(0.009)	(0.009)	(0.008)	(0.008)
Country Europe	0.003	0.003	−0.004	0.003	0.001	0.005	0.003	0.001	0.004	0.003
	(0.015)	(0.014)	(0.015)	(0.015)	(0.014)	(0.013)	(0.013)	(0.013)	(0.011)	(0.012)
Country America	0.024	0.020	0.009	0.022	0.017	0.020	0.021	0.019	0.022	0.023
	(0.020)	(0.021)	(0.025)	(0.025)	(0.021)	(0.020)	(0.019)	(0.020)	(0.019)	(0.018)
Vendor 2			−0.063**							
			(0.023)							
Vendor 3			−0.104**							
			(0.035)							
Vendor 4			0.001							
			(0.037)							
Vendor 5			0.021							
			(0.019)							
Vendor 6			−0.146***							
			(0.028)							
Vendor 7			−0.018							
			(0.016)							
Vendor 8			−0.023							
			(0.032)							
Vendor 9			−0.078*							
			(0.037)							
Vendor 10			−0.084**							
			(0.032)							
Vendor 11			−0.101**							
			(0.036)							
Vendor 12			−0.017							
			(0.025)							
Vendor 13			−0.018							
			(0.027)							
Vendor 14			−0.023							
			(0.031)							
Vendor 15			−0.056							
			(0.186)							
Number of segments		−0.005		−0.005	−0.005	−0.005	−0.005	−0.005	−0.005	−0.005
		(0.006)		(0.006)	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)	(0.005)
Reach (users profiled)		−0.012		−0.013	−0.013	−0.012	−0.012	−0.013^{+}	−0.013^{+}	−0.013^{+}
		(0.008)		(0.009)	(0.008)	(0.008)	(0.008)	(0.007)	(0.007)	(0.007)
Online Shopping	0.012	0.012	0.012	0.010	0.035^{+}			0.034*		
	(0.022)	(0.021)	(0.026)	(0.018)	(0.021)			(0.015)		

Table 9. (Continued)

Dependent variable:	*Likelihood of being profiled accurately* (mean DV = 0.341, n = 30,513)									
	(1) Logit	(2) Logit	(3) Logit	(4) Logit	(5) Logit	(6) Logit	(7) Logit	(8) OLS	(9) OLS	(10) OLS
Online activities	0.006	0.006	0.006	0.006		0.009^{+}			0.009*	
	(0.006)	(0.006)	(0.006)	(0.006)		(0.005)			(0.004)	
Devices	0.009*	0.009*	0.009*	0.008*			0.012***			0.012***
	(0.004)	(0.004)	(0.004)	(0.004)			(0.003)			(0.003)
(McFadden) R^2	0.218	0.218	0.220	0.219	0.214	0.216	0.217	0.271	0.273	0.274
Log-likelihood	−15,326.1	−15,315.6	−15,280.3	−15,303.2	−15,396.6	−15,350.8	−15,339.5			
AIC	30,698.2	30,681.2	30,634.6	30,674.3	30,839.2	30,747.7	30,725.0	31,459.1	31,366.8	31,345.4
BIC	30,889.7	30,889.3	30,942.6	30,957.4	31,030.7	30,939.1	30,916.5	31,659.0	31,566.6	31,545.2

Notes. All models have clustered standard errors (by individual, attribute, country). FE, fixed effects; DV, dependent variable. OLS clustered standard errors are based on wild bootstraps (10,000 draws).
$^{+}p < 0.1$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

side variables are related and provide a theoretical account for our findings. Previous research on digital inequality has demonstrated differences in internet behaviors between socio-economically advantaged and disadvantaged users who have more or less access to online technologies (DiMaggio et al. 2004, Ono and Zavodny 2007, Hsieh et al. 2008). Hence, in line with digital-divide literature, consumers with a less privileged socio-economic status should have a smaller digital footprint (i.e., be in data deserts) that hinder vendors from properly profiling these consumers.

7.1. Method and Model

To examine the relationship of digital-footprint and consumer characteristics for our context, we went back to the panel provider and were able to obtain relevant measures on both digital-behavior and socio-economic variables for 1,267 and 1,547 panelists for our home-country and international sample, respectively. We then investigated the relationship of users' background characteristics and their digital footprint by regressing the three relevant characteristics—online shopping, electronic devices owned, and regular online activities—on our socio-economic variables. We used a logit model for the binary outcome of online shopping and Poisson regressions for the count variables of how many devices a user owns and how many regular online activities in which someone engages.[16] For the international sample, we also include model specifications again with two-way interactions of country fixed effects and our two social-economic variables (log(*Income*) and *Female*). To provide cluster-robust variance estimation for the international sample with few clusters for our logit and Poisson models, we apply the bias-reduced linearization procedure developed by Pustejovsky and Tipton (2017).

7.2. Results

Table 11 shows the results of nine regressions for our two samples. We note that the models without interaction terms (columns (4), (6), and (8)) show the better fit to our data for the international sample, according to both the AIC and the BIC, even though the average marginal effects are, again, barely affected by the decision on whether to include the interactions in our models.

In line with our previous findings, we find associations with different coefficient signs for being female, suggesting a mixed overall relationship between gender and digital footprint. That is, we can see that women are more likely to shop online, even though they tend to possess fewer electronic devices and engage in fewer regular online activities. We remind the reader that our online activity measure relates to the number of diverse activities in which someone engages (= breadth of activities), and not the intensity of individual activities.[17]

Table 10. Average Marginal Effects: Digital-Behavior Characteristics and Online Coverage

Dependent variable:	*Likelihood of being profiled across data brokers* (mean DV = 0.310, n = 61,665)									
	(1) Logit	(2) Logit	(3) Logit	(4) Logit	(5) Logit	(6) Logit	(7) Logit	(8) OLS	(9) OLS	(10) OLS
Country Asia-Pacific	−0.0233***	−0.0233***	−0.0232***	−0.0220***	−0.0229***	−0.0228***	−0.0231***	−0.0230***	−0.0228***	−0.0233***
	(0.0005)	(0.0005)	(0.0003)	(0.0000)	(0.0002)	(0.0001)	(0.0002)	(0.0005)	(0.0006)	(0.0006)
Country Europe	0.0143***	0.0143***	0.0143***	0.0140***	0.0142***	0.0145***	0.0145***	0.0141***	0.0145***	0.0145***
	(0.0003)	(0.0003)	(0.0003)	(0.0000)	(0.0003)	(0.0001)	(0.0000)	(0.0007)	(0.0002)	(0.0004)
Country America	0.2471***	0.2471***	0.2476***	0.2486***	0.2463***	0.2467***	0.2477***	0.2464***	0.2467***	0.2471***
Country interactions				Yes						
Vendor 2			0.0559***	0.0559***	0.0559***	0.0559***	0.0559***	0.0559***	0.0559***	0.0559***
			(0.0139)	(0.0139)	(0.0140)	(0.0140)	(0.0139)	(0.0139)	(0.0138)	(0.0139)
Vendor 3			0.9336***	0.9336***	0.9336***	0.9336***	0.9336***	0.9336***	0.9336***	0.9336***
			(0.0226)	(0.0225)	(0.0226)	(0.0226)	(0.0226)	(0.0489)	(0.0491)	(0.0491)
Vendor 4			−0.0107	−0.0107	−0.0107	−0.0107	−0.0107	−0.0107	−0.0107	−0.0107
			(0.0076)	(0.0076)	(0.0076)	(0.0076)	(0.0076)	(0.0118)	(0.0119)	(0.0119)
Vendor 5			0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
			(0.0107)	(0.0107)	(0.0107)	(0.0107)	(0.0107)	(0.0093)	(0.0092)	(0.0093)
Vendor 6			−0.0253	−0.0253	−0.0253	−0.0253	−0.0253	−0.0253	−0.0253	−0.0253
			(0.0518)	(0.0518)	(0.0518)	(0.0518)	(0.0518)	(0.0579)	(0.0583)	(0.0582)
Vendor 7			0.0039	0.0039	0.0039	0.0039	0.0039	0.0039	0.0039	0.0039
			(0.0064)	(0.0064)	(0.0064)	(0.0064)	(0.0064)	(0.0038)	(0.0038)	(0.0038)
Vendor 8			0.0671***	0.0671***	0.0671***	0.0671***	0.0671***	0.0671***	0.0671***	0.0671***
			(0.0150)	(0.0150)	(0.0150)	(0.0150)	(0.0150)	(0.0182)	(0.0181)	(0.0183)
Vendor 9			0.9353***	0.9353***	0.9353***	0.9353***	0.9353***	0.9353***	0.9353***	0.9353***
			(0.0223)	(0.0223)	(0.0223)	(0.0223)	(0.0223)	(0.0489)	(0.0490)	(0.0491)
Vendor 10			0.7850***	0.7850***	0.7850***	0.7850***	0.7850***	0.7850***	0.7850***	0.7850***
			(0.0677)	(0.0671)	(0.0677)	(0.0677)	(0.0677)	(0.0714)	(0.0717)	(0.0719)
Vendor 11			0.9385***	0.9385***	0.9385***	0.9385***	0.9385***	0.9385***	0.9385***	0.9385***
			(0.0215)	(0.0214)	(0.0215)	(0.0215)	(0.0215)	(0.0482)	(0.0484)	(0.0484)
Vendor 12			0.0032***	0.0032***	0.0032***	0.0032***	0.0032***	0.0032^{+}	0.0032*	0.0032^{+}
			(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0016)	(0.0016)	(0.0017)
Vendor 13			0.2313^{+}	0.2313^{+}	0.2313^{+}	0.2313^{+}	0.2313^{+}	0.2313*	0.2313*	0.2313*
			(0.1322)	(0.1321)	(0.1322)	(0.1322)	(0.1322)	(0.1050)	(0.1036)	(0.1047)
Vendor 14			−0.0182^{+}	−0.0182^{+}	−0.0182^{+}	−0.0182^{+}	−0.0182^{+}	−0.0182	−0.0182	−0.0182
			(0.0094)	(0.0094)	(0.0094)	(0.0094)	(0.0094)	(0.0173)	(0.0173)	(0.0173)
Vendor 15			−0.0525*	−0.0525*	−0.0525*	−0.0525*	−0.0525*	−0.0525	−0.0525	−0.0525
			(0.0223)	(0.0222)	(0.0223)	(0.0223)	(0.0223)	(0.0494)	(0.0496)	(0.0496)
Number of segments		0.1511***								
		(0.0130)								
Reach (users profiled)		0.0148								
		(0.0307)								
Online Shopping	0.0024	0.0023	0.0026	0.0052***	0.0048			0.0045		
	(0.0042)	(0.0042)	(0.0042)	(0.0001)	(0.0045)			(0.0043)		
Online activities	−0.0003	−0.0003	−0.0004	−0.0004***		0.0002			0.0002	
	(0.0011)	(0.0011)	(0.0010)	(0.0000)		(0.0006)			(0.0005)	

Table 10. (Continued)

Dependent variable:	*Likelihood of being profiled across data brokers* (mean DV = 0.310, n = 61,665)									
	(1) Logit	(2) Logit	(3) Logit	(4) Logit	(5) Logit	(6) Logit	(7) Logit	(8) OLS	(9) OLS	(10) OLS
Devices	0.0018	0.0018	0.0017	0.0016***			0.0016*			0.0017^{+}
	(0.0014)	(0.0014)	(0.0011)	(0.0000)			(0.0008)			(0.0010)
(McFadden) R^2	0.019	0.107	0.705	0.706	0.705	0.705	0.705	0.748	0.748	0.748
Log-likelihood	−37,467.0	−34,110.3	−11,255.7	−11,235.5	−11,263.6	−11,264.5	−11,256.4			
AIC	74,947.9	68,238.5	22,553.4	22,531.0	22,565.1	22,566.9	22,550.9	−5,126.7	−5,125.2	−5,140.9
BIC	75,011.1	68,319.8	22,743.0	22,801.9	22,736.7	22,738.5	22,722.4	−4,946.2	−4,944.6	−4,960.3

Notes. All models have clustered standard errors (by individual and country). FE, fixed effects; DV, dependent variable. OLS clustered standard errors are based on wild bootstraps (10,000 draws).
$^{+}p < 0.1$; $^{*}p < 0.05$; $^{***}p < 0.001$.

Most importantly, in support of our main research focus that income is crucial for profiling accuracy and coverage, we find that individuals with a higher income are more likely to shop online, to engage in a greater number of online activities, and to own many electronic devices. This positive association reaches statistical significance for eight of our nine models.

In addition, we find that people who have a college degree are more likely to engage in a larger number of regular online activities, whereas those with blue-collar jobs tend to engage in fewer online activities (column (3)). Also, those in family households seem to have more electronic devices and are more likely to shop online, whereas those in multilingual households tend to have fewer devices (column (2)), but engage in more online activities.

Our results largely support the findings of previous digital-divide studies and illustrate how the socio-economic background of individuals can be linked to internet behavior and the ability to leave a digital footprint that allows data brokers to build profiles. Some individuals have the means and time to be online more often, thus creating rich pools of data. Others have a limited online presence and thus represent data deserts, whose poor signals make successful profiling more challenging.

8. Discussion and Conclusion

Existing studies have documented a large degree of heterogeneity in successful consumer profiling (Neumann et al. 2019, Venkatadri et al. 2019, De Bruyn and Otter 2022). Our work examines the relationship of firm- and consumer-side characteristics and third-party audience profiling to shed light on what factors may explain differences in profiling. Using five field studies covering 15 data brokers, we investigate both profiling accuracy and the availability of profiling attributes for an individual (coverage) for demographic attributes.

We present evidence that both profiling outcomes (coverage and accuracy) are linked to socio-economic status. Profiles are more accurate for individuals from affluent households, individuals with college degrees, family households, and multilingual households. People with blue-collar jobs or racial minorities are less likely to be profiled correctly. The combined influence of socio-economic background variables can make a difference in profiling accuracy of up to 18.7%, which is large, given that overall accuracy was 36.6%, on average, for the online marketing campaigns we studied.[18]

The role of socio-economic variables is similar, but less pronounced, for profiling coverage. Which consumer variables correlate with higher coverage depends on whether this the context is online or offline or the geography in which an online campaign is running. We find that consumers with college degrees tend to be

Table 11. Average Marginal Effects: Socio-Economic Characteristics and Digital-Behavior Outcomes

Sample:	Home country (n = 1,267)			International (n = 1,547)					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent variable:	*Online Shopping*	*Devices*	*Online Activities*	*Online Shopping*	*Online Shopping*	*Devices*	*Devices*	*Online Activities*	*Online Activities*
Model:	Logit	Poisson	Poisson	Logit	Logit	Poisson	Poisson	Poisson	Poisson
Country interactions					Yes		Yes		Yes
Country Europe				0.0187***	0.0200***	0.2605**	0.1885***	−0.1496***	−0.1034***
				(0.0007)	(0.0015)	(0.0973)	(0.0171)	(0.0216)	(0.0024)
Country America				−0.0134***	−0.0134***	−0.1562**	−0.2863***	−0.0672***	0.0140**
				(0.0010)	(0.0020)	(0.0493)	(0.0169)	(0.0079)	(0.0053)
log(income)	0.0189*	0.3143***	0.0904	0.0153***	0.0152***	0.6857***	0.6826***	0.2818***	0.2815***
	(0.0094)	(0.0936)	(0.0684)	(0.0010)	(0.0007)	(0.1032)	(0.0119)	(0.0267)	(0.0061)
Female	0.0485**	−0.5999***	−0.5814***	0.0245***	0.0247***	−0.3843***	−0.3824***	−0.4013***	−0.4061***
	(0.0187)	(0.1787)	(0.1303)	(0.0041)	(0.0015)	(0.0269)	(0.0165)	(0.0443)	(0.0095)
Blue–collar job	0.0071	−0.2453	−0.5403**						
	(0.0264)	(0.2582)	(0.2012)						
Shared HH	0.0238	0.8158**	0.2480						
	(0.0238)	(0.2592)	(0.1879)						
Family HH	0.0497*	1.2106***	0.3733*						
	(0.0219)	(0.2349)	(0.1701)						
College degree	0.0143	0.2062	0.9534***						
	(0.0190)	(0.1701)	(0.1224)						
Multilingual HH	−0.0103	−0.3716^{+}	0.3248*						
	(0.0229)	(0.2186)	(0.1517)						
Mean DV	0.90	3.95	7.21	0.94	0.94	3.63	3.63	7.01	7.01
Log.Lik.	−390.6	−2,826.3	−2,985.6	−350.1	−349.9	−3,438.2	−3,436.3	−3,565.2	−3,564.7
AIC	797.2	5,668.5	5,987.2	710.3	717.9	6,886.4	6,890.6	7,140.3	7,147.4
BIC	838.4	5,709.7	6,028.4	737.0	766.0	6,913.1	6,938.7	7,167.1	7,195.5

Notes. International-sample models have clustered standard errors (by country). FE, fixed effects; HH, household; DV, dependent variable; Log.Lik., log-likelihood.
$^{+}p < 0.1$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

profiled online more often, whereas individuals from multilingual households are profiled less often. We also find that offline coverage, again, seems to depend on the ethnic and racial background, with minorities being profiled less often. In line with our profiling accuracy results, the consumer variable that appears most important in coverage analyses is a consumer's affluence (income or poverty).

What could explain the critical role of a person's socio-economic background—in particular, income—for successful profiling? One possible explanation may be that data brokers focus specifically on lucrative consumers for a marketing context. However, it is important to note that data brokers have no economic incentive to only focus on a subset of consumers, as the business model is purely volume-driven (the profiles are sold on a fixed-price basis per thousands). The more profile attributes they can offer to advertisers, the more revenue that brokers can generate. Furthermore, advertisers buy the demographic data to identify consumers with a certain age, gender, or family status because they think there is a good product match (e.g., targeting women to sell hygiene products like tampons). It would not be appropriate to only have high-income earners if advertisers did not specify this criterion.

Instead of supply-side factors, we present evidence that the digital divide (DiMaggio et al. 2004) may underpin our findings: people's socio-economic characteristics determine the degree of digital footprint that they leave online, which, in turn, determines profiling accuracy and coverage. We show that higher-income people are more likely to own many electronic devices, shop online, and engage regularly in many diverse online activities, such as news sharing or video gaming. The breadth of their measurable digital signals creates data pools valuable to data brokers as inputs for consumer profiles. In contrast, blue-collar or lower-income individuals or those living alone do not provide as many online signals, leading to data deserts that impede successful profiling.

Profiling coverage strongly depends on differences across individual brokers for profiles that marketers can access in online campaigns using data-management platforms, explaining 65.9%–84.6% of variation. For example, the number of profile segments and, depending on the examined market, the number of users profiled correlate with profiling coverage. Speculatively, other firm-side factors may be important, too, such as partnerships that allow access to certain mobile apps, websites, or offline data partners.

Profiling accuracy, by contrast, does not appear to strongly depend on the underlying characteristics of a data broker, accounting for 0.1%–0.4% of variation in our context. This finding suggests that data brokers may not differ much in their profiling methods, such as the way they use machine learning applications to infer profile attributes. The ability to access a large number of users seems to have no impact on the accuracy of consumer profiles. This implies that learning about the correct feature of our examined demographic attributes cannot be achieved through large data volumes alone. However, there still may be an indirect effect of data-broker characteristics on profiling accuracy. Because data brokers differ in their coverage, they differ in their access to people online, who, therefore, come from different socio-economic backgrounds and are profiled to different degrees of accuracy.

8.1. Implications

Our findings suggest that marketers should carefully consider what customers they try to reach when buying demographic audience segments from third-party data brokers for ad campaigns. Some population groups, such as single households or less affluent consumers, are less likely to be profiled correctly and are, hence, harder to address when relying on segment attributes for audience targeting. To reach less affluent consumers, organizations should seek other methods of identification rather than prebuilt audience segments.

Our results also suggest that marketers should be wary of targeting methods that rely on identifier matching to avoid pitfalls related to identity fragmentation bias (Lin and Misra 2022). If we compare the accuracy of age attributes offline versus online for our data, we find an average accuracy of 84.2% for our offline exercise versus 19.0%–23.3% for the online data (see columns for *Age* in Table 2). Although this is only suggestive evidence,[19] this stark difference implies that a large proportion of errors may not be due to incorrect profile attributes created by brokers, but, rather, due to other errors in the process of providing profiles online, such as matching profiles to the wrong cookies.

Regarding policy implications, the influence of background characteristics leading to different data availability and accuracy should be of concern to policy makers. Many of the data brokers supplying marketing platforms with data also provide demographic information for credit decisions, insurance decisions, and other types of background checks and risk assessments. Our research provides the first empirical evidence that both available and accurate digital profiles depend on who you are—with strong differences between the "poor" and the "rich." This finding is particularly problematic because it is very hard for individuals to obtain detailed information about one's own profile attributes that data brokers created, making it cumbersome to correct wrong information about a profile attribute (Miller 2017).

Our findings do support one potential policy shift. The Interactive Bureau of Advertising (IAB), an industry body, proposed introducing a data-transparency label, as depicted in Figure 3, that lists essential information

Figure 3. (Color online) The Data Transparency Standard Label Proposed by the IAB

Data Transparency Facts

Data Distributor Name: **Data Company**
Data Distributor Contact: DataSolutionTeam@data.com
Data Provider Name: **Leasing Company**
Data Provider Contact: DataAccounts@leasingco.com

Audience Snapshot	
Branded Name	Auto Intenders – Six Months
Standard Name	Auto Intenders
Audience Description	Households likely in the market to purchase a new vehicle in the next six months
Geographies	USA

Audience Construction	Attributes
Audience Count	6,500,000
Precision Level	Households
Activation ID(s)	Cookies
Audience Expansion	Yes
Cross-Device Expansion	Yes
Last Refresh Date	02-Jan-2018
Event Lookback Window	60 Days

Data Source	Attributes
Source ID Description	Dealer-reported names and postal codes of individuals who requested test drives
Source ID Contribution	1,130,000
Precision Level	Individual
ID Key	Name and Postal
Source Event	Transactions
Inclusion Method	Observed
Seed Size (if modeled)	-
Source Refresh Frequency	Quarterly
Event Lookback Window	180 Days

This Data Transparency Label has been developed by members of ANA's Council for Data Integrity and IAB Tech Lab's Data Transparency Working Group, with the support of CIMM, The ARF and IAB's Data Center of Excellence. For more information, please visit datalabel.org.

about the data sources and methods that were used to create segments in 2018 (IAB 2018). Because adding the proposed data label is voluntary, almost no vendor has adopted it as of the writing of this article. Regulators should consider making such data labels mandatory, similar to the laws that were introduced in the food industry for nutrition labels.

The finding that profiles of large households and affluent households are more likely accurately reconstructed is important for privacy considerations. If we view privacy regulation as effectively dismantling the ability of data brokers to use black-box approaches to profile people, this suggests that privacy protection may have the largest effective consequences for affluent consumers and family households. Regulators should carefully consider how further policies can be created to address privacy protection for these population groups.

8.2. Limitations and Future Research Opportunities

As with all research, our study is subject to limitations, which often represent opportunities for promising future research, too. First, our results are based on the most common demographic attribute types and 15 data brokers. Future studies should examine other audience attributes and contexts under the lens of profiling errors and their drivers. Second, our analysis on profiling success drivers is based on observational data and associations, due to the nature of the research problem we examine. We, however, believe that the most likely explanation for our consistent patterns across our studies, such as the relationship of high income and profiling success, is causal. Third, although we show a link between digital footprint and socio-economic status, in Web Appendix A.6, we show that digital consumer behavior alone is not the only variable contributing to the role of people's background that matters for profiling accuracy and coverage.[20] Another possible source of profiling errors can be related to algorithmic bias and the lack of diverse training samples (Lambrecht and Tucker 2019). Although this is hard to tackle without insider access, we believe this is a fruitful area for future research. Finally, we find that profile prices in our regressions did not have explanatory power. Analyzing the role of pricing for profile attributes, which was beyond the scope of this work, is another useful avenue for future research efforts.

Given the lack of transparency and the ongoing importance of the global consumer profiling industry, it is crucial to expand the current literature and understand what drivers may affect profiling quality. Our study provides novel insights into factors influencing profiling outcomes for marketing, with potential implications for other areas (credit and risk decisions). We hope that our work will provide analytical guidance for future studies and fuel further discussions of how to improve both general profiling practices and legislation to increase transparency for consumers.

Acknowledgments

The authors thank Lingguo Xu for the research assistance. All errors are their own.

Endnotes

[1] See, for example, the 2014 Federal Trade Commission (FTC) report on data brokers at https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf.

[2] Four weeks is the typical industry standard used for most web cookies and also a common period for online campaigns.

[3] ISO 20252: 2019 Market, Opinion and Social Research.

[4] *State of residence* refers to the federal state in which an individual resides.

[5] Age was reported in different tiers, such as 18–25, 20–29, 35–44, 30–39, etc. We treated each tier as an individual profile attribute. A detailed overview of the different age tiers can be found in Web Appendix A.2. Likewise, location data comprise information on five regions, each of which we treat as individual attribute.

[6] The CPM is indicated in U.S. dollars in all four countries.

[7] Wild bootstrapping only works for linear regression (Roodman and Morduch 2014, Roodman et al. 2019).

[8] We thank an anonymous reviewer for this suggestion.

[9] Although not shown here, the difference in R^2 for OLS models is of the same magnitude.

[10] The panel uses the same income buckets and number of possible tiers for each country.

[11] Source of data: https://dl.ncsbe.gov/index.html?prefix=data/.

[12] It should be noted that "Hispanic" and "non-Hispanic" are descriptions of ethnicity, whereas "Asian," "Black," and "white" are descriptions of race. Our "Hispanic" list contains Hispanic individuals of any race, whereas we refer to the "non-Hispanic" category lists by their race to keep labels concise.

[13] Poverty itself is defined by the Office of Management and Budget's (OMB's) Statistical Policy Directive 14; see https://www.census.gov/topics/income-poverty/poverty/guidance/poverty-measures.html.

[14] Similar to income, ZIP-code poverty shows better fit when being log-transformed.

[15] This is a different subset of all panelists and only partially overlaps with the sample from Studies 1 and 2.

[16] For example, for the outcome of number of devices owned by person i given covariates X_i, the Poisson model assumes $Prob(Devices_i|X_i) = \frac{\lambda_i^{Devices_i}}{Devices_i!} exp^{-\lambda_i}$, and $\lambda_i = e^{(X_i\beta)}$.

[17] In line with our findings, social media user statistics confirm that men seem to be more represented on most social media platforms (Statista 2023).

[18] This calculation is based on blue-collar versus white-collar jobs, college versus non-college education, the type of household, and a 100% difference in income. The average accuracy number is based on home-country market, which was slightly higher than the international sample, which had an average of 33.5% overall.

[19] We only report the overall average of brokers, as we cannot reveal which of the 15 brokers in Study 1 was the same broker as in Study 3. However, if we compare the numbers just for this broker, there is a stark difference between online and offline accuracy.

[20] Likewise, our finding that minorities are less often profiled successfully seems not completely explainable by ZIP-code poverty proportions (see Figure A2 in Web Appendix A.4).

References

Adjerid I, de Matos MG (2019) Consumer consent and firm targeting after GDPR: The case of a large telecom provider. Mimeo, Virginia Tech, Blacksburg, VA.

Ansari A, Mela C (2003) E-customization. *J. Marketing Res.* 40(2):131–145.

Aziz A, Telang R (2016) What is a cookie worth? Preprint, submitted April 2, https://dx.doi.org/10.2139/ssrn.2757325.

Binns R, Lyngs U, Van Kleek M, Zhao J, Libert T, Shadbolt N (2018) Third party tracking in the mobile ecosystem. *Proc. 10th ACM Conf. Web Sci.* (Association for Computing Machinery, New York), 23–31.

Cameron AC, Miller DL (2015) A practitioner's guide to cluster-robust inference. *J. Human Resources* 50(2):317–372.

Cameron AC, Gelbach JB, Miller DL (2008) Bootstrap-based improvements for inference with clustered errors. *Rev. Econom. Statist.* 90(3):414–427.

De Bruyn A, Otter T (2022) Bayesian consumer profiling: How to estimate consumer characteristics from aggregate data. *J. Marketing Res.* 59(4):755–774.

DiMaggio P, Hargittai E, Celeste C, Shafer S (2004) From unequal access to differentiated use: A literature review and agenda for research on digital inequality. Neckerman K, ed. *Social Inequality* (Russell Sage Foundation, New York), 355–400.

Flosi S, Fulgoni G, Vollman A (2013) If an advertisement runs online and no one sees it, is it still an ad? Empirical generalizations in digital advertising. *J. Adverting Res.* 53(2):192–199.

FTC (2014) Data brokers: A call for transparency and accountability. Report, Federal Trade Commission, Washington, DC.

Gartner (2020) Information technology glossary: Data broker. Accessed January 27, 2023, https://www.gartner.com/en/information-technology/glossary/data-broker.

Goldberg S, Johnson G, Shriver S (2019) Regulating privacy online: The early impact of the GDPR on European web traffic & e-commerce outcomes. Preprint, submitted July 17, https://dx.doi.org/10.2139/ssrn.3421731.

Goldfarb A, Prince J (2008) Internet adoption and usage patterns are different: Implications for the digital divide. *Inform. Econom. Policy* 20(1):2–15.

Goldfarb A, Tucker C (2012) Privacy and innovation. *Innovation Policy Econom.* 12(1):65–90.

Heckman JJ (1979) Sample selection bias as a specification error. *Econometrica* 47(1):153–161.

Hoffman DL, Novak TP, Schlosser A (2000) The evolution of the digital divide: How gaps in Internet access may impact electronic commerce. *J. Comput.-Mediated Commun.* 5(3):JCMC534.

Hsieh JP-A, Rai A, Keil M (2008) Understanding digital inequality: Comparing continued use behavioral models of the socio-economically advantaged and disadvantaged. *MIS Quart.* 32(1):97–126.

IAB (2018) Major advertising trade bodies unveil data transparency label. Accessed January 27, 2023, https://iabtechlab.com/standards/the-data-transparency-standard-data-label/.

IAB, WinterberryGroup (2020) The state of data. Accessed January 10, 2021, https://www.iab.com/wp-content/uploads/2020/07/IAB-Winterberry_Group_The_State_of_Data_2020_July_2020.pdf.

Jia J, Jin GZ, Wagman L (2018) The short-run effects of GDPR on technology venture investment. NBER Working Paper 25248, National Bureau of Economic Research, Cambridge, MA.

Johnson G, Shriver S (2019) Privacy & market concentration: Intended & unintended consequences of the GDPR. Preprint, submitted November 15, https://dx.doi.org/10.2139/ssrn.3477686.

Johnson GA, Shriver SK, Du S (2020) Consumer privacy choice in online advertising: Who opts out and at what cost to industry? *Marketing Sci.* 39(1):33–51.

Johnson GA, Shriver SK, Goldberg SG (2023) Privacy and market concentration: Intended and unintended consequences of the GDPR. *Management Sci.* 69(10):5695–5721.

Keller J (1995) Public access issues: An introduction. Kahin B, Keller J, eds. *Public Access to the Internet* (MIT Press, Cambridge, MA), 34–45.

Lambrecht A, Tucker C (2019) Algorithmic bias? An empirical study into apparent gender-based discrimination in the display of STEM career ads. *Management Sci.* 65(7):2966–2981.

Lee H, Wong SF, Oh J, Chang Y (2019) Information privacy concerns and demographic characteristics: Data from a Korean media panel survey. *Government Inform. Quart.* 36(2):294–303.

Levine MM (1995) A brief history of information brokering. *Bull. Amer. Soc. Inform. Sci.* 21:8–9.

Lewis RA, Reiley DH (2014) Online ads and offline sales: Measuring the effect of retail advertising via a controlled experiment on Yahoo! *Quant. Marketing Econom.* 12(3):235–266.

Li Y (2022) Cross-cultural privacy differences. Knijnenburg BP, Page X, Wisniewski P, Lipford HR, Proferes N, Romano J, eds. *Modern Socio-Technical Perspectives on Privacy* (Springer International Publishing, Cham, Switzerland), 267–292.

Lin T, Misra S (2022) Frontiers: The identity fragmentation bias. *Marketing Sci.* 41(3):433–440.

Lucker J, Hogan S, Bischoff T (2017) Predictably inaccurate: The prevalence and perils of bad big data. *Deloitte Rev.* 21:8–25.

Manchanda P, Dubé J-P, Goh KY, Chintagunta PK (2006) The effect of banner advertising on internet purchasing. *J. Marketing Res.* 43(1):98–108.

Miller CR (2017) I bought a report on everything that's known about me online. *Atlantic* (June 6), https://www.theatlantic.com/technology/archive/2017/06/online-data-brokers/529281/.

Miller AR, Tucker CE (2011) Can healthcare information technology save babies? *J. Polit. Econom.* 119(2):289–324.

Murthi BPS, Sarkar S (2003) The role of the management sciences in research on personalization. *Management Sci.* 49(10):1344–1362.

Neumann N, Tucker CE, Whitfield T (2019) Frontiers: How effective is third-party consumer profiling? Evidence from field studies. *Marketing Sci.* 38(6):918–926.

Ono H, Zavodny M (2007) Digital inequality: A five country comparison using microdata. *Soc. Sci. Res.* 36(3):1135–1155.

Petrison LA, Blattberg RC, Wang P (1997) Database marketing: Past, present, and future. *J. Direct Marketing* 11(4):109–125.

Pustejovsky JE, Tipton E (2017) Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *J. Bus. Econom. Statist.* 36(4):672–683.

Roodman D, Morduch J (2014) The impact of microcredit on the poor in Bangladesh: Revisiting the evidence. *J. Dev. Stud.* 50(4):583–604.

Roodman D, Nielsen MØ, MacKinnon JG, Webb MD (2019) Fast and wild: Bootstrap inference in Stata using boottest. *Stata J.* 19(1):4–60.

Servon LJ (2008) *Bridging the Digital Divide: Technology, Community and Public Policy* (John Wiley & Sons, Hoboken, NJ).

Statista (2023) Gender distribution of social media audiences worldwide as of January 2022, by platform. Accessed January 27, 2023, https://www.statista.com/statistics/274828/gender-distribution-of-active-social-media-users-worldwide-by-platform/.

Stiebellehner S, Wang J, Yuan S (2017) Learning continuous user representations through hybrid filtering with doc2vec. Preprint, submitted December 31, https://arxiv.org/abs/1801.00215.

The Economist (2017) Fuel of the future: Data is giving rise to a new economy. *Economist* (May 6), https://www.economist.com/briefing/2017/05/06/data-is-giving-rise-to-a-new-economy.

TMR (2022) Data brokers market: Global industry analysis, size, share, growth, trends and forecast, 2021–2031. Accessed January 27, 2023, https://www.transparencymarketresearch.com/data-brokers-market.html.

Trusov M, Ma L, Jamal Z (2016) Crumbs of the cookie: User profiling in customer-base analysis and behavioral targeting. *Marketing Sci.* 35(3):405–426.

Tucker CE, Yu S (2019) Does IT lead to more equal treatment? An empirical study of the effect of smartphone use on customer complaint resolution. Preprint, submitted June 12, https://dx.doi.org/10.2139/ssrn.3011633.

Van de Ven WP, Van Praag BM (1981) The demand for deductibles in private health insurance: A probit model with sample selection. *J. Econometrics* 17(2):229–252.

Venkatadri G, Sapiezynski P, Redmiles EM, Mislove A, Goga O, Mazurek M, Gummadi KP (2019) Auditing offline data brokers via Facebook's advertising platform. Liu L, White R, eds. *WWW'19 World Wide Web Conf.* (Association for Computing Machinery, New York), 1920–1930.

Wooldridge J (2010) *Econometric Analysis of Cross Section and Panel Data* (MIT Press, Cambridge, MA).

Zukowski T, Brown I (2007) Examining the influence of demographic factors on Internet users' information privacy concerns. *SAICSIT'07 Proc. 2007 Annu. Res. Conf. South African Inst. Comput. Sci. Inform. Tech. IT Res. Developing Countries* (Association for Computing Machinery, New York), 197–204.